INRIA
CENTRE DE ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. (3) 954 90 20

Rapports de Recherche

N° 447

L'ÉTUDE DES TABLEAUX *n*-AIRES PAR LA CLASSIFICATION AUTOMATIQUE

Henri RALAMBONDRAINY

Octobre 1985

L'ETUDE DES TABLEAUX n-AIRES PAR LA CLASSIFICATION AUTOMATIQUE

Henri RALAMBONDRAINY
INRIA
Domaine de Voluceau
BP 105 - Rocquencourt
78153 Le Chesnay Cedex (France)

RESUME

L'Analyse des Données est considérée comme un ensemble de techniques de réduction de l'information. Nous définissons la mesure de l'information d'un tableau comme la norme du tenseur associé au tableau et étudions la réduction de cette information. L'extension de cette définition et des techniques de réduction de l'information aux tableaux n-aires fournit des généralisations des méthodes factorielles et de classifications type Nuées dynamiques.

ABSTRACT

The Data Analysis methods are considered as algorithms for information reduction. We define the data array measure of information as the norm of the tensor associated to the data array and study the reduction of this information. The generalization of this definition and the algorithms of information reduction give generalization of factorial analysis and clustering methods as Nuées Dynamiques.

SOMMAIRE

I - INTRODUCTION

I.1 Les données

Si l'on examine les différents axes de recherche actuels en Analyse de Données (A. D.), on s'aperçoit que l'étude de suite de tableaux, tableau n-aire par exemple, tiennent une grande importance. Cela vient du fait que les techniques d'A. D. concernant un seul tableau de données sont désormais classiques et que les utilisateurs sont de plus en plus confrontés à des tableaux multiples. Ce sont par exemple des suites de tableaux indicés par le temps ou bien des groupes de variables mesurées sur le même ensemble d'individus etc, ... Diverses méthodes ont été proposées dont nous allons rappeler rapidement les principales caractéristiques.

I.2 Les principales approches

Supposons que nous avons un tableau X(I, J, T) indicé par le temps t $\in$ T une première possibilité pour analyser de tels tableaux et de se ramener à des tableaux à double entrées en considérant le tableau $X(I, J_T)$ où les tableaux sont juxtaposés en colonnes, tableau appelé par Benzécri comme le tableau "histoire des variables" ou celui où les tableaux sont juxtaposés en lignes : $X(I_T, J)$ tableau appelé 'histoire des individus". Il est alors possible d'appliquer les techniques d'A. D. classiques. Lorsque l'on a un ensemble de variables mesurées sur un même ensemble d'individus, Caroll [Car 70] a proposé l'Analyse Canonique Généralisée (A. C. G.) qui recherche des variables combinaisons linéaires des groupes de variables, les plus liées au sens de la liaison R^2 coefficient de corrélation multiple aux différents groupes.

Si les groupes de variables sont de type qualitatif, la méthode est l'analyse des Correspondances Multiples (A.C.M.). Lorsque les variables sont de types quantitatifs, la méthode A.C.G. est peu utilisée, car elle pose des problèmes d'interprétation. La variance des groupes expliquée par les variables canoniques peut en effet être faible. Faisant ces remarques, Escofier - Pages [EsP 84] proposent alors une autre mesure de liaison notée L^2 fournissant des variables canoniques décrivant mieux les groupes au sens de la variance expliquée. Cette liaison L^2 entre une variable v normée et un groupe de variables X_q est définie comme l'inertie de l'ensemble de projections sur v des variables

du groupe. Se fondant sur une telle liaison, ces auteurs proposent alors la méthode l'Analyse Factorielle Multiple qui correspond à une Analyse en Composantes Principales pondérée, permettant des représentations conjointes des variables, groupes de variables et individus.

Dans une autre optique, Escoufier [Esc 80] a proposé une autre approche pour étudier globalement un ensemble de tableaux. L'idée est d'associer un opérateur U_q à chaque groupe de variables dont les vecteurs propres sont les composantes principales du tableau, puis de définir une proximité entre tableaux mesurés par une distance entre opérateurs. Ainsi deux tableaux seront considérés comme prôches ou semblables, si leurs nuages de variables ont même forme. On pourra par exemple se référer à notre étude de la charge d'un ordinateur [Ral 79] où nous analysons un tableau de données évoluant dans le temps. Nous avons alors appliqué et comparé diverses méthodes factorielles, de classification automatique et la méthode des opérateurs. Des développements plus récents ont été apporté concernant cette approche [Lhr 76] [Fou 84]. Elle consiste à l'étude de "l'interstructure" c'est-à-dire le positionnement multidimensionnel des opérateurs, l'étude de l'opérateur compromis, c'est-à-dire de l'opérateur résumant au mieux les différents opérateurs puis l'étude de l'intra-structure c'est-à-dire la représentation graphique des variables et individus de l'étude.

Le point commun entre ces différentes approches est l'optique "factorielle" choisie par ces auteurs. C'est-à-dire que la recherche de représentations graphiques optimales des individus ou variables ont une grande importance, car c'est le moyen choisi pour appréhender l'information relative aux données. Notre optique est tout autre, se situant dans le cadre de la classification automatique et nous n'aborderons pas de tels problèmes de visualisation de données.

Nous partirons d'une considération générale que l'Analyse des données a pour objectif, entre autre, d'extraire l'information pertinente des données. Nous définirons ensuite ce que nous entendons par mesure d'information associée aux données, puis étudierons de manière formelle les techniques de décomposition optimale de cette mesure d'information et enfin nous proposerons divers algorithmes dans le cadre de la classification automatique de réduction de la mesure d'information.

I.3 Philosophie et concepts généraux

Nous considérons les méthodes d'Analyse des Données comme des techniques de réduction de l'information relative aux données. On peut montrer en effet qu'une famille de méthodes factorielles. (Analyse en Composantes Principales, Analyse des Correspondances par exemple) et des méthodes de Classification Automatique (Nuées Dynamiques : méthode des centres de gravités par exemple) fournissent des réductions optimales de l'information relative à un tableau de données, mesure d'information qui est l'inertie du nuage des individus à l'occurence. Le type de structure axes ou points utilisés simplement diffère. Govaert [Gov 83] adopte et propose alors la démarche suivante pour étudier un ensemble de tableaux types : tableau de mesures, contingence, binaire et modalités. Pour chaque type de tableau X, il définit une mesure d'information I(X) et propose divers algorithmes permettant d'avoir une bonne approximation de I(X) par la recherche simultanée de partitions P et Q sur l'ensemble d'individus et variables tel que l'information I(P, Q) soit une réduction optimale de I(X).

Nous suivrons cette démarche pour l'étude de tableaux n-aires. Nous aurons donc les étapes suivantes :

1) Définition d'une mesure d'information relative à un ensemble de Q tableaux X_q. Dans le cas de tableau unique, il est assez facile de définir une mesure d'information d'un tel tableau. Lorsque l'on a un ensemble de tableaux hétérogènes, le problème est plus complexe. Nous proposons de considérer comme mesure d'information relative à un tableau n-aire, la moyenne pondérée des mesures d'informations relatives à chaque tableau. Cette idée naturelle reporte la difficulté sur les choix des coefficients de pondérations. Divers choix sont possibles, nous les présentons et en discutons.

2) Proposer diverses techniques pour une réduction optimale de la mesure d'information ainsi définie. Les méthodes proposées sont alors un ensemble d'algorithmes résolvant des problèmes d'optimisation relatives à la décomposition de la mesure d'information. Suivant le type de structure adopté, nous obtiendrons des méthodes factorielles ou de classification automatique généralisant les techniques classiques.

Traditionnellement les méthodes traitant des tableaux n-aires partent de situations précises. Ainsi les méthodes type analyse de correspondances multiples et leurs variantes (analyse de tableaux de Burt, pondérations des questions, etc, ...) développés par Cazes [Caz 80] par exemple, concernent des variables

qualitatives avec des métriques du chi-deux. L'Analyse Factorielle Multiple proposée par Escofier, traite des tableaux quantitatifs ou qualitatifs mais impose que les métriques relatives aux individus soient diagonales. Les constructions faites sont en général, dès le départ sous-tendues par des objectifs de traitement des données, les préoccupations d'interprétation des résultats ayant une grande importance. Il nous semble toutefois que du point de vue théorique, il est préférable de dégager le problème mathématique commun à ce type d'approche, de le traiter dans toute sa généralité et ensuite de voire les applications des résultats dans le contexte précis de l'Analyse de Données. Ainsi nous nous affranchissons dans un premier temps de diverses contraintes usuelles en Analyse des Données : par exemple que la métrique relative aux variables est diagonale, que lorsque l'on traite des données quantitatives, le nuage des individus est centré, etc, ... Nous démontrons le maximum de résultats possibles et ensuite nous les interprétons dans les cadres de l'Analyse des Données.

I.4 Cadres et outils mathématiques :

L'objet d'étude de l'Analyse des Données est un tableau X qui peut être considéré comme une application linéaire de E^* dans F où E est l'espace d'individus muni d'une métrique M et F, l'espace des variables étant munies d'une métrique N que nous supposerons quelconque. En Analyse des Données on s'intéresse à $X \in \mathcal{L}(E^*,F)$ si l'on étudie l'ensemble des individus et à $X' \in \mathcal{L}(F^*,E)$ si l'on s'intéresse aux variables. Du point de vue mathématique, les deux situations sont symétriques et il est préférable de considérer le tenseur $X_{E \otimes F}$ de l'espace $E \otimes F$ car X et X' sont l'expression de ce même tenseur. Les espaces E et F étant munis des métriques M et N, l'espace produit tensoriel $E \otimes F$ sera muni de la métrique produit tensoriel $M \otimes N$. On donnera comme définition de mesure d'information du tableau X dans le contexte des métriques M, N, la norme du tenseur $X_{E \otimes F}$ dans $E \otimes F$. La mesure d'information de X est donc $I(X) = \| X_{E \otimes F} \|^2_{M \otimes N}$, on montre qu'elle généralise la notion usuelle d'inertie.

D'autres cadres de références que $(E \otimes F, M \otimes N)$ seront envisagés. Ainsi à un triplet (X, M, N) Escoufier [Esc 80] propose d'associer des opérations $U \in L(F,F)$ et $Z \in L(E',E)$ respectivement N-symétrique et M-symétrique caractéristiques du triplet (X, M, N) et dont les éléments principaux ont un grande importance en Analyse des Données.

Les vecteurs propres de U sont en effet les composantes principales du triplet (X, M, N). En vertu des isomorphismes $L(F, F) \simeq F^* \otimes F$ et $L(E, E) \simeq E^* \otimes E$ nous utiliserons la représentation tensorielle des opérateurs U et Z qui seront considérés comme des tenseurs mixtes $U \in F^* \otimes F$ et $Z \in E^* \otimes E$. Ces espaces produits tensoriels étant munis du produit scalaire canonique trace. Nous considèrerons donc aussi pour l'étude de la réduction de l'information I(X) les espaces produits tensoriels $(F^* \otimes F, < >)$ et $(E^* \otimes E, < >)$. Des liens existent entre ces tenseurs de références $X_{E \otimes F}$, U et Z qui seront mis en évidence et la mesure d'information I(X) sera interprétée dans les différents cadres choisis.

Pour généraliser les définitions et propriétés précédentes à l'étude de tableaux n-aires, nous supposons que l'on a un ensemble de triplets $(X_{\alpha\beta}, M_\alpha, N_\beta)$ où $\alpha \in A$ et $\beta \in B$, A et B étant des ensembles finis. Nous avons donc un ensemble de couples, espaces, métriques : (E_α, M_α), (F_β, N_β), $\alpha \in A$, $\beta \in B$ correspondant aux espaces relatifs à $X_{\alpha\beta}$. On considère alors les espaces sommes orthogonales $E = \Sigma\{E_\alpha \mid \alpha \in A\}$, $F = \Sigma\{F_\beta \mid \beta \in B\}$ et les métriques pondérées $M = \Sigma\{c_\alpha M_\alpha \mid \alpha \in A\}$ et $N = \Sigma\{c'_\beta N_\beta \mid \beta \in B\}$ relatives à E et F où $\{c_\alpha \mid \alpha \in A\}$ et $\{c'_\beta \mid \beta \in B\}$ sont des coefficients positifs. On considère ensuite l'espace produit tensoriel $E \otimes F$ qui est la somme orthogonale des espaces $E_\alpha \otimes F_\beta$, $\alpha \in A$, $\beta \in B$, il est muni de la métrique produit $M \otimes N$, la mesure d'information relative aux triplets est alors la norme $\| X_{E\otimes F} \|^2_{M\otimes N}$ de $X_{E\otimes F} = \pi\{X_{E_\alpha \otimes F_\beta} \mid \alpha \in A, \beta \in B\}$, elle est la moyenne pondérée des mesures d'information $\| X_{E_\alpha \otimes F_\beta} \|^2_{M_\alpha \otimes N_\beta}$ relatives aux triplets $(X_{\alpha\beta}, M_\alpha, N_\beta)$. Nous étudions alors les différents problèmes d'optimisation relatifs à la réduction de cette information dans les différents cadres de référence considérés c'est-à-dire $(E \otimes F, M \otimes N)$, $(F^* \otimes F, < >)$ et $(E^* \otimes E, < >)$.

Il apparaît donc que les objets et espaces manipulés seront essentiellement des tenseurs et des produits tensoriels d'espaces vectoriels. Les raisons essentielles de ces choix sont outre la puissance du calcul tensoriel, la possibilité d'interpréter géométriquement les propriétés relatives à un ensemble de tableaux, en effet un tableau est alors représenté par un tenseur donc un point dans un espace vectoriel. On se reportera par exemple aux diverses réflexions de Benzecri dans [Ben 73] sur l'insuffisance du calcul matriciel et la nécessité de l'approche tensorielle des problèmes d'Analyse des Données. En particulier le problème de la réduction de l'information de $I(X) = \| X_{E\otimes F} \|^2_{M\otimes N}$

dans ($E \otimes F$, $M \otimes N$) apparaît comme la recherche de l'approximation d'ordre K d'un tenseur. Présentation que donne Benzecri de l'analyse factorielle dans [Ben 73] sur la leçon concernant la réduction d'un élément du produit tensoriel de deux espaces euclidiens.

C'est partant de ce point de vue que nous abordons l'étude des tableaux n-aires. Nous supposerons connues les principales propriétés du produit tensoriel dont on pourra avoir de plus amples précisions dans [Sch 81].

I.5 Applications à l'analyse des données

Nous l'envisagerons sous plusieurs points de vue :

1) Du point de vue des méthodes factorielles :

Les problèmes d'optimisation présentés généralisent ceux relatifs à l'analyse des correspondances multiples ou l'analyse factorielle multiple et d'une manière générale ceux relatifs à l'analyse pondérée d'un ensemble de tableaux par les méthodes factorielles. Nous justifions et éclairons donc de telles pratiques.

Si nous considérons un tableau n-aire comme un ensemble de groupe de variables mesurées sur un même ensemble d'individus, l'approche proposée revient à chercher un ensemble de variables v liées au sens d'une liaison aux différents groupes de variables. Cette liaison $\mathcal{L}$ qui s'exprime comme la somme pondérée des corrélations au carré d'un vecteur v et des facteurs relatifs à chaque groupe de variables, est la forme la plus générale de liaison dans ce type d'approche. On montre en effet que les liaisons R^2 de Caroll et L^2 d'Escofier ne sont que des expressions de pour des métriques particulières. La principale difficulté dans ces méthodes réside dans la détermination des coefficients de pondérations, nous présentons les principales possibilités en les éclairant dans notre contexte.

La mise en oeuvre des techniques d'Analyse de Données nécessite un certain nombre de choix de base : choix de l'ensemble d'individus, choix de l'ensemble des variables, choix des métriques pour les variables et l'ensemble des individus (pondérations "intra"). L'approche proposée pour l'étude de tableaux n-aires ajoute donc un nouveau choix de base aux précédents qui est le choix relatif aux pondérations "inter" c'est-à-dire un ensemble de coefficients positifs pour équilibrer les rôles joués par les différents groupes de variables.

2) Du point de vue de la classification automatique

Les méthodes de classification que nous envisageons ici sont celles relatives aux Nuées Dynamiques (appelée M. N. D.). La caractéristique principale de telles méthodes, c'est la recherche de recouvrement optimal au sens d'un critère exprimant l'adéquation entre un recouvrement de l'ensemble des objets à classer et un mode de représentation des classes de ce recouvrement (cf [Did 79]). Les types de recouvrement que nous traitons sont les partitions et suivant le mode de représentation choisi, on aboutit à des problèmes d'optimisations et des algorithmes différents. Les résultats théoriques précédents permettent alors de généraliser un ensemble de méthodes de type Nuées Dynamiques à l'étude de tableau n-aire, considéré comme un ensemble d'individus mesurés par des paquets de variables pouvant être de natures diverses. On propose alors les méthodes suivantes :

- Nuées Dynamiques Généralisées :

Le mode de représentation choisi est le centre de gravité, le critère optimisé est l'inertie-inter, nous interprétons le critère sous diverses formes suivant les cadres de références choisis. En particulier, nous montrons que la MND. généralisée sur un paquet de variables $\{X_q \mid q \in Q\}$ revient à chercher une variable qualitative X_k la plus liée au sens de $\mathcal{L}$ aux variables X_q.

- L'Analyse Factorielle Typologique Généralisée :

L'Analyse Factorielle Typologique a été proposée par OK [OK 75] pour généraliser les méthodes factorielles classiques. Ainsi au lieu de chercher la variété affine de dimension q la plus proche du nuage à analyser, on cherche k-variétés affines de dimension q les plus proches du nuage. Nous l'étendons à l'étude de paquets de variables, en particulier lorsque toutes les variables sont de type qualitatives, nous avons une généralisation de l'Analyse des Correspondances Multiples.

- L'Analyse Canonique Typologique Généralisée :

Sous le nom de l'Analyse Canonique Typologique, Diday [Did 78] propose un certain nombre d'algorithmes permettant de détecter des liaisons locales entre variables suivant leurs types (quantitatifs ou qualitatifs). En effet, les techniques classiques telles que l'Analyse Canonique ne tiennent pas compte de l'hétérogénéité de la population étudiée pour exhiber des liens entre variables ou combinaisons linéaires de variables. Le cas que nous étudions ici est le cas général où l'on étudie plus de deux paquets de variables pour détecter des liaisons locales.

De même que l'analyse canonique peut être considérée comme une Analyse en Composantes Principales particulières, nous étudions les liens existants entre l'Analyse Factorielle Typologique Généralisée et l'Analyse Canonique Typologique Généralisée.

Dans un certain sens, les méthodes factorielles (Analyse Factorielle Multiple, STATIS) sont complémentaires des méthodes proposées car les techniques graphiques qu'elles proposent peuvent être ensuite appliquées sur les différentes classes obtenues par ces techniques de partitionnement.

Dans une première partie nous examinons les approches existantes : la méthode des opérateurs, l'Analyse Canonique Généralisée de Caroll et l'Analyse Factorielle Multiple d'Escofier, puis nous présentons les différents concepts et cadres de référence relatifs à notre approche et nous appliquerons enfin les différents résultats obtenus à la classification automatique.

II LES PRINCIPALES APPROCHES

II.1 Les données et notations

Nous supposons que nous avons Q groupes de variables $\{X_q \mid q \in Q\}$ mesurées sur un même ensemble d'individus I. On note comme il est habituel l'ensemble et son cardinal par la même lettre. Ainsi :

I désigne l'ensemble des individus et son cardinal card I = n

J_q désigne l'ensemble des indices des variables rélatifs au groupe q et le nombre de variables de ce groupe

$J = U \{J_q \mid q \in Q\}$ désigne l'ensemble total des indices des variables et le nombre total de variables : card J = p

$X_q = \begin{bmatrix} 1 & \cdots j \cdots & J_q \\ & \vdots & 1 \\ & x_i^j & \vdots \\ & & i \\ & & \vdots \\ & & I \end{bmatrix}$ désigne le tableau des variables relatives au groupe q, $\underline{x}_{iq}$ désignera l'individu i de ce groupe $\underline{x}^j$, $j \in J_q$, la variable j de ce groupe

$E_q \simeq R^{J_q}$ l'espace des individus relatif au groupe q

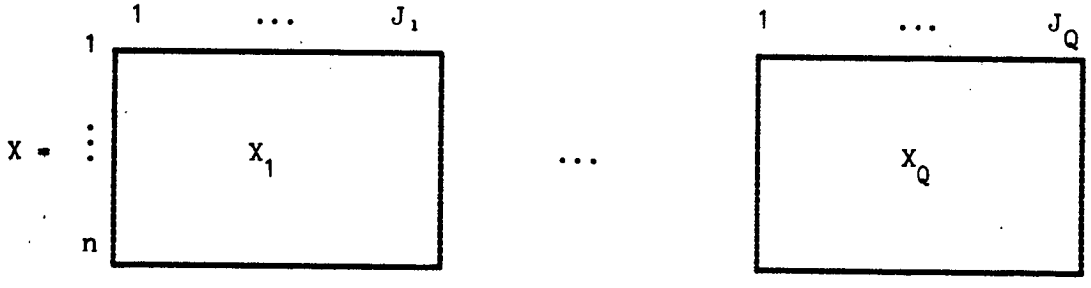
M_q la métrique associée au tableau q et aussi l'application duale de E_q dans E_q^* et la matrice $J_q \times J_q$ correspondante

$F \simeq R^I$ désigne l'espace de référence des variables

D_p l'ensemble des pondérations relatives aux individus définissant une métrique sur F

II.2 L'espace des individus $E \simeq \mathbb{R}^J$

Pour représenter globalement les individus relativement à l'ensemble des variables, on considère l'espace $E = \mathbb{R}^J$ somme orthogonale directe des espaces $E_q = \mathbb{R}^{J_q}$: on a donc $E = \Sigma\{E_q \mid q \in Q\}$ ou $\mathbb{R}^J = \Sigma\{\mathbb{R}^{J_q} \mid q \in Q\}$ dont la dimension est $J = \Sigma\{J_q \mid q \in Q\}$. Dans cet espace, un individu $\underline{x}_i$ a Q composantes $\underline{x}_{iq} \in E_q$, on note $\underline{x}_i = \pi\{\underline{x}_{iq} \mid q \in Q\}$. Il correspond à la ligne i du tableau X juxtaposition des tableaux X_q : $X = \pi\{X_q \mid q \in Q\}$.



On note π_q la projection canonique de E sur E_q et i_q l'injection canonique de E_q dans E.

$\pi_q : E \longrightarrow E_q$ $\qquad$ $i_q : E_q \longrightarrow E$

$\underline{x}_i \longrightarrow \pi_q(\underline{x}_i) = \underline{x}_{iq}$ $\qquad$ $\underline{x}_{iq} \longrightarrow i_q(\underline{x}_{iq}) = \tilde{\underline{x}}_{iq}$

où $\tilde{\underline{x}}_{iq}$ a tous ses sous vecteurs $\tilde{\underline{x}}_{iq'}$ nuls sauf la qème égale à $\underline{x}_{iq}$. Nous avons le nuage des individus N_E^I ou N_J^I relatif à l'espace total E et Q nuages relatifs aux espaces E_q que l'on peut plonger dans E

$$N_{Jq}^I \text{ ou } N_{Eq}^I = \{i_q\ (\underline{x}_{iq}),\ i \in I\} \subset E$$

Soit $\underline{g}$ le centre de gravité du nuage N_E^I : $\underline{g} = \Sigma \{p_i\ \underline{x}_i \mid i \in I\}$ on a :

$$\pi_q\ (\underline{g}) = \Sigma \{P_i\ \pi_q\ (\underline{x}_i) \mid i \in I\}$$

$$\underline{g}_q = \Sigma \{P_i\ \underline{x}_{iq} \mid i \in I\}$$

la composante $\underline{g}_{iq} \in E_q$ est donc le centre de gravité du nuage $N_{E_q}^I$ dans E_q. On définit un produit scalaire M sur E de la manière suivante :

Proposition

Le produit scalaire M pondéré par les coefficients positifs $\{c_q \mid q \in Q\}$ est défini sur E comme suit :

soient $\underline{x} = \pi \{\underline{x}_q \mid q \in Q\}$ et $\underline{y} = \pi \{\underline{y}_q \mid q \in Q\}$; $\underline{x}, \underline{y} \in E$.

$$M\ (\underline{x}, \underline{y}) = c_1\ M_1\ (\underline{x}_1, \underline{y}_1) + \ldots + c_q\ M_Q\ (\underline{x}_Q, \underline{y}_Q)$$

Il est facile de vérifier que M est un produit scalaire car les coefficients $\{c_q \mid q \in Q\}$ sont positifs et les M_q sont des produits scalaires. Matriciellement M est une matrice diagonale par blocs.

$$M = \begin{array}{c|c|c|c} & 1 \ldots J_1 & & 1 \ldots J_Q \\ \hline & C_1\ M_1 & 0 & 0 \\ \hline & 0 & \ddots & 0 \\ \hline & 0 & 0 & C_Q\ M_Q \end{array}$$

et si $\underline{x}_i$ et $\underline{x}_{i'}$ sont deux individus de E l'on a :

$$d^2_M (\underline{x}_i, \underline{x}_{i'}) = \Sigma \{c_q \, d^2_{M_q} (\underline{x}_{iq}, \underline{x}_{i'q}) \mid q \in Q\}$$

La distance entre deux individus est donc la moyenne pondérée des distances entre ces composantes.

Comme $J = U \{J_q \mid q \in Q\}$, on définit alors l'application : a de J dans Q

$a : J \longrightarrow Q$

$j \longrightarrow a(j) = q$ où q est l'ensemble auquel appartient j. La matrice peut alors s'écrire :

$$M = \{M^{jj'} = \delta^{a(j)}_{a(j')} \, c_{a(j)} \, M^{jj'}_{a(j)} \mid j,j' \in J\} \quad \text{(II.2.1)}$$

où δ désigne le symbolde de Kronecker.

II.3 L'espace des opérateurs $F^* \otimes F \simeq \mathbf{R}^{I*} \otimes \mathbf{R}^I$ [Cap 76], [Esc 80], [Bra 73]

Au triplet (X_q, M_q, D_p) est associé le schéma de dualité de l'ACP.

$$\begin{array}{ccc} E_q \simeq R^{J_q} & \xrightarrow{X'_q} & F^* \simeq R^{I*} \\ M_q \downarrow \uparrow V_q & & W_q \downarrow \uparrow Dp \\ E^*_q \simeq R^{J_q*} & \xrightarrow{X_q} & F \simeq R^I \end{array}$$

Pour étudier globalement les liaisons entre les groupes de variables X_q, nous allons placer dans un espace de dimension card I x card I qui sera :

$L(\mathbf{R}^I, \mathbf{R}^I)$: l'espace des applications linéaires de $\mathbf{R}^I$ dans $\mathbf{R}^I$

ou $\mathbf{R}^{I*} \otimes \mathbf{R}^I$: l'espace des tenseurs mixtes isomorphe à $L(\mathbf{R}^I, \mathbf{R}^I)$.

Dans cet espace chaque triplet (X_q, M_q, D_p) sera représenté par un vecteur unique :

$U_q = W_q D_p$ l'opérateur d'Escoufier D_p-symétrique de $L(\mathbf{R}^I, \mathbf{R}^I)$ ou tenseur D_p-symétrique de $\mathbf{R}^{I*} \otimes \mathbf{R}^I$.

Cet opérateur est représentatif du triplet (X_q, M_q, D_p) (cf [Cal 76]) car uniquement fonction des distances entre individus

$$D^2_{i_q i'_q} = \| \underline{x}_{i_q} - \underline{x}_{i'_q} \|^2_{M_q}$$

Les composantes principales associées au tableau X_q ne sont autres que les vecteurs propres de l'opérateur U_q.

II.3.1 Produit scalaire dans l'espace $\mathbf{R}^{I*} \otimes_I \mathbf{R}^I$

Nous utiliserons de préférence la représentation tensorielle des opérateurs en nous plaçant dans l'espace $\mathbf{R}^{I*} \otimes \mathbf{R}^I$. En effet les aspects géométriques des problèmes sont mieux mis en évidence et les calculs facilités.

Rappelons que le produit tensoriel de rang 1 entre une forme linéaire $\alpha \in \mathbf{R}^{I*}$ et un vecteur $u \in \mathbf{R}^I$ est l'application linéaire $\alpha \otimes u \in L(\mathbf{R}^I, \mathbf{R}^I)$ tel que :

$\forall x \in \mathbf{R}^I \quad \alpha \otimes u(x) = \alpha(x).u$

Si $\{f_i \mid i \in I\}$ est une base de $\mathbf{R}^I$ et $\{f_i^* \mid i \in I\}$ la base duale de $\mathbf{R}^{I*}$ alors $\{f_i^* \otimes f_j \mid i, j \in I\}$ forment une base de $\mathbf{R}^{I*} \otimes \mathbf{R}^I$.

Un tenseur $T \in \mathbf{R}^{I*} \otimes \mathbf{R}^I$ s'exprime dans cette base :

$$T = \Sigma \{T_{ij}\ f_i^* \otimes f_j \mid i, j \in I\}$$

si S est un autre tenseur de $\mathbb{R}^{I*} \otimes \mathbb{R}^I$

$$S = \Sigma \{S_{kl}\ f_k^* \otimes f_l \mid k, l \in I\}$$

Par définition la trace du tenseur $T \in \mathbb{R}^{I*} \otimes \mathbb{R}^I$ notée trace T est le nombre unique indépendant de la représentation T définie en remplaçant chaque produit tensoriel $f_i^* \otimes f_j$ par la "contraction" $\langle f_i^*, f_j \rangle = f_i^*(f_j)$

$$\text{trace } T = \Sigma \{T_{ij}\ f_i^*(f_j) \mid i, j \in I\} = \Sigma \{T_{ii} \mid i \in I\}$$

on retrouve la définition de la trace relative à une matrice.

L'espace $\mathbb{R}^{I*} \otimes \mathbb{R}^I$ est alors muni du produit scalaire canonique :

$$\langle T, S \rangle = \text{trace } (T.S) = \text{trace } T.S$$

En particulier pour u un vecteur de $\mathbb{R}^I$, on note u^* la forme linéaire $D_p(u) \in \mathbb{R}^{I*}$ où D_p est considérée comme l'isomorphisme entre $\mathbb{R}^I$ et $\mathbb{R}^{I*}$ induit par la métrique D_p, cad. : $u^* = D_p(u) = \langle ., u \rangle_{D_p}$ alors par la définition le produit scalaire entre $u^* \otimes v$ et $u'^* \otimes v'$ $\forall (u, v, u', v') \in \mathbb{R}^I$ s'écrit :

$$\langle u^* \otimes v, u'^* \otimes v' \rangle = \text{trace } (u^* \otimes v \circ u'^* \otimes v') = \langle u, v' \rangle_{D_p} \langle u', v \rangle_{D_p}$$

en utilisant la propriété de "contraction" du produit tensoriel.

Ainsi si U_q et $U_{q'}$ sont les tenseurs associés aux triplets (X_q, M_q, D_p) et $(X_{q'}, M_{q'}, D_{p'})$, le produit scalaire de U_q et $U_{q'}$ dans $\mathbb{R}^{I*} \otimes \mathbb{R}^I$ a pour expression en utilisant leurs expressions matricielles :

$$\boxed{\langle U_q, U_{q'} \rangle = \text{trace } (W_q\ D_p\ W_q\ D_p)}$$

II.3.2 Propriétés des opérateurs

L'opérateur U_q étant D_p-symétrique est diagonalisable. On considère $\{\phi_i^q \mid i \in I\}$ la base D_p-orthonormée formée des vecteurs propres de U_q, X associés à des valeurs propres non nulles complétés par une base du noyau de U_q, de $\mathbf{R}^I$. Ainsi $\{\phi_i^{q*} \otimes \phi_j^q \mid i,j \in I\}$ est une base de $\mathbf{R}^{I*} \otimes \mathbf{R}^I$ dans laquelle la matrice représentative de U_q est diagonale, les éléments diagonaux étant les valeurs propres $\{\lambda_i^q \mid i \in I\}$ de U_q. On a alors une représentation tensorielle de U_q :

$$U_q = \Sigma \{\lambda_i^q \, \phi_i^{q*} \otimes \phi_i^q \mid i \in I\}$$

De même si $\{\phi_j^{q'} \mid j \in I\}$ sont les vecteurs propres unitaires de $U_{q'}$, associés aux valeurs propres de $\{\lambda_j^{q'} \mid j \in I\}$ alors :

$$U_{q'} = \Sigma \{\lambda_j^{q'} \, \phi_j^{q'*} \otimes \phi_j^{q'} \mid j \in I\}$$

En utilisant les propriétés de linéarité de l'opérateur trace et la contraction du produit tensoriel :

$$\boxed{\langle U_q, U_{q'} \rangle = \Sigma \{\lambda_i^q \, \lambda_j^{q'} (\langle \phi_i^q, \phi_j^{q'} \rangle)^2 \mid i, j \in I\}}$$

On en tire alors les résultats suivants :

Propriétés :

i) La norme de l'opérateur U_q et la distance entre U_q et $U_{q'}$ s'écrivent :

$$\| U_q \|^2 = \Sigma \{\lambda_i^{q^2} \mid i \in I\}$$

$$\| U_q - U_{q'} \|^2 = \Sigma \{\lambda_i^{q^2} \mid i \in I\} + \Sigma \{\lambda_j^{q'^2} \mid j \in I\}$$

$$- 2 \Sigma \{\lambda_i^q \, \lambda_j^{q'} (\langle \phi_i^q, \phi_j^{q'} \rangle_{D_p})^2 \mid i, j \in I\}$$

ii) Une condition nécessaire et suffisante pour que deux opérateurs soient égaux qu'ils aient même spectre (valeurs propres identiques et vecteurs propres en bijection).

On considère donc que les triplets (X_q, M_q, D_p) et $(X_{q'}, M_{q'}, D_p)$ sont équivalent si les opérateurs U_q et $U_{q'}$ sont identiques ou proportionnels. En adoptant comme type de proximité entre tableaux la distance entre opérateurs on compare les tableaux selon la forme de leurs nuages variables ou individus. Deux tableaux étant considérés comme proches si les nuages associés ont mêmes directions d'allongement en vertu de l'expression (II.3.2.1.).

II.3.3 Liaison entre triplets

On définira la liaison entre deux triplets (X_q, M_q, D_p) et $(X_{q'}, M_{q'}, D_p)$ comme le produit scalaire entre les tenseurs U_q et $U_{q'}$ de $\mathbb{R}^{I*} \otimes \mathbb{R}^{I}$. On note $\mathcal{L}((X_q, M_q, D_q), (X_{q'}, M_{q'}, D_q)) = \langle U_q, U_{q'} \rangle$ ou plus simplement en omettant la métrique D_p commune :

$$\mathcal{L}((X_q, M_q), (X_{q'}, M_{q'})) = \langle U_q, U_{q'} \rangle$$

en normalisant par les normes $\| U_{q'} \|$, on a l'équivalent d'un coefficient de corrélation R_V (cf[Esc 80]) qui est égale à 1 si U_q et $U_{q'}$ sont proportionnels c'est-à-dire les deux triplets sont équivalents. Toutefois, nous l'utiliserons sous la forme non normalisée.

Suivant le type de tableau et les métriques considérées, la liaison s'interprètera différemment. Nous allons donner quelques unes de ces expressions sans démonstration et on se reportera à [Caï 76] pour plus de détails.

- Si on associe à un tableau de variables quantitatives X_q la métrique de Mahalanobis V_{qq}^{-1}, l'opérateur U_q s'écrit :

$$U_q = X_q V_{qq}^{-1} X_q' D_p = A_q \text{ le } D_p\text{-projecteur sur l'espace } X_q' (E_q^*)$$

Soit deux triplets (X_q, V_{qq}^{-1}, D_p), $(X_{q'}, V_{q'q'}^{-1}, D_p)$, on a alors :

$$\mathcal{L}((X_q, V_{qq}^{-1}), (X_{q'}, V_{q'q'}^{-1})) = \langle A_q, A_{q'} \rangle = \Sigma \{r_k^2 \mid k \in K\}$$

où r_k est le $k^{\text{ème}}$ coefficient de corrélation canonique entre X_q et $X_{q'}$ et $k = \inf (J_q, J_{q'})$.

- Si on considère des triplets relatifs à des variables qualitatives $(X_q, D_{1|P_q}, D_p)$ et $(X_{q'}, D_{1|Pq'}, D_p)$, les métriques associées étant celles du chi-deux $D_{1|Pq} = (X'_q D_p X_q)^{-1}$ les opérateurs associés sont les projecteurs A_q et $A_{q'}$ sur X'_q (E^*_q) et $X'_{q'}$ $(E^*_{q'})$ et l'on a :

$$\mathcal{L}((X_q, D_p), (X_{q'}, D_p)) = < A_q, A_{q'} > = \phi^2_{qq'} + 1$$

où $\phi^2_{qq'}$ est le phi-deux associé au tableau P à J_q lignes et $J_{q'}$ colonnes des probabilités $p_{jj'}$ d'association des modalités de J_q et $J_{q'}$ et l'on a

$$\phi^2_{qq'} = \frac{X^2 (J_q, J_{q'})}{n}$$

où $X^2 (X_q, X_{q'})$ est le chi-deux associé au tableau de contingence croisant les variables X_q et $X_{q'}$ et n=card I le nombre d'individus.

- Soient les triplets (X_q, V^{-1}_{qq}, D_p), $(X_{q'}, D_{1|Pq'}, D_p)$ associés à un tableau quantitatif X_q et une variable qualitative X_q, on a alors :

$$\mathcal{L}((X_q, V^{-1}_{qq}), (X_{q'}, D_{1|Pq'})) = < A_q, A_{q'} > = \text{trace } (V^{-1}_{qq} B) = Ir\ (N^G_{E_q})$$ où $B = G\ D_p\ G'$ est la matrice d'inertie "inter-classe", G désignant le tableau des vecteurs centres de gravité relatifs aux modalités de X_q, trace $(V^{-1}_{qq} B)$ est alors l'inertie des centres de gravité du nuage $N^G_{E_q}$ dans E_q.

II.3.4 Pratique des opérateurs

Nous allons présenter rapidement l'approche proposée par Escoufier pour l'étude conjointe de plusieurs matrices de données quantitatives. Les principes sont celles des méthodes factorielles, cad l'accent est mis sur la représentation visuelle des opérateurs, individus et variables à partir du tableau de distances ou produits scalaires entre opérateurs (on aurait pu imaginer un traitement par la classification automatique du tableau des distances entre opérateurs). Cette approche consiste à l'étude de :

l'interstructure :

Chaque tableau étant représenté par son opérateur, on cherche le positionnement multidimensionnel optimal des Q opérateurs à partir de la matrice des produits scalaires entre opérateurs. Cela revient à faire une analyse factorielle sur le tableau des distances entre opérateurs.

le compromis :

La deuxième étape de la méthode est la recherche d'un opérateur résumant au mieux l'ensemble des opérateurs c'est l'opérateur compromis qui est une combinaison linéaire des opérateurs U_q

$$T' = \Sigma \{c_q U_q \mid q \in Q\}$$

où $\{C_q \mid q \in Q\}$ est le vecteur propre normé ($\Sigma \{c_q \mid q \in Q\} = 1$) associé à la plus grande valeur λ propre de l'analyse précédente

les intra-structures :

Il est intéressant pour l'utilisateur d'avoir une représentation des variables et individus, diverses techniques complexes sont proposées suivant le type d'opérateurs considérés.

Pour plus de détails on se reportera aux références suivantes : [Lhr 76], [Fou 84] et pour une comparaison de ces différentes techniques à [Gla 81].

Nous n'allons pas toutefois adopter cette approche, mais retenir le cadre conceptuel c.a.d l'idée d'associer à un triplet, des opérateurs représentatifs dont nous allons donner les expressions générales.

II.3.5 Expression générale des tenseurs associés à un triplet

Nous avons déjà donné une représentation tensorielle des opérateurs en fonction de leurs éléments propres. Dans ce paragraphe, nous en fournissons d'autres qui seront utilisées ultérieurement. Soient les tenseurs :

$U \in F^* \otimes F$ et $Z \in E^* \otimes E$ associés à un triplet (X, M, N) où M et N sont des métriques quelconques sur E et F, dont les matrices sont : $M = \{M^{jj'} \mid j, j' \in J\}$ et $N = \{N_{ii'} \mid i, i' \in I\}$. L'opérateur U a pour expression

$$U = WN = XMX'N = \Sigma \{M^{jj'} \underline{x}^{j'} \underline{x}'^{j} N \mid j, j' \in J\}$$

par suite $\forall$ $z \in F$, l'image de z par U est :

$$Uz = \Sigma \{M^{jj'} \underline{x}^{j'} \underline{x}'^{j} Nz \mid j, j' \in J\}$$

$$= \Sigma \{M^{jj'} N (z, \underline{x}^{j}) \underline{x}^{j'} \mid j, j' \in J\}$$

d'où une expression tensorielle de U :

$$\boxed{U = \Sigma \{M^{jj'} \underline{x}^{j*} \otimes \underline{x}^{j'} \mid j,j' \in J\}} \qquad (II.3.5.1)$$

De manière symétrique, le tenseur $Z \in E^* \otimes E$ a pour expression :

$$\boxed{Z = \Sigma \{N_{ii'} \underline{x}_{i'} \otimes \underline{x}_{i} \mid i,i' \in I\}} \qquad (II.3.5.2)$$

II.4 L'analyse canonique généralisée de Caroll : la liaison R^2 [Cac70]

II.4.1 Introduction

Pour étudier un ensemble Q de variables X_q, Caroll propose la recherche d'une variable $v \in R^I$, la plus liée à l'ensemble des variables au sens suivant ; v est solution du problème :

Problème II.4.0

$$\left| \begin{array}{l} \max \Sigma \{R^2 (v, X_q) \mid q \in Q\} \\ v \in R^I \end{array} \right. \qquad \text{(II.4.1)}$$

où $R^2 (v, X_q)$ est la corrélation multiple entre v et le groupe de variables $\{\underline{x}^j \mid j \in J_q\}$.

II.4.2 Présentation

L'approche de Caroll revient donc à chercher un vecteur $v \in R^I$, solution du problème.

Problème II.4.1

$$\left| \begin{array}{l} \max \Sigma \{< v, A_q v >_{D_p} \mid q \in Q\} \\ v \in R^I \\ \text{avec } \| v \|^2_{D_p} = 1 \end{array} \right.$$

où A_q désigne le D_p-projecteur sur l'espace W_q engendré par les variables relatives à X_q. Le problème (II.4.1) s'écrit alors :

Problème II.4.1

$$\left| \begin{array}{l} \max \quad <v, (\Sigma \{A_q \mid q \in Q\}) v >_{D_p} \\ v \in R^I \\ \text{avec } \| v \|^2_{D_p} = 1 \end{array} \right.$$

Problème classique de maximisation du quotient de deux formes quadratique dont la solution est rappelée par la proposition.

Proposition II.4.1

La variable v la plus liée à l'ensemble de variables X_q, au sens de liaison R^2, est le vecteur propre de $A_Q = \Sigma \{A_q \mid q \in Q\}$ associée à la plus grande valeur propre λ.

Les variables canoniques ξ_q sont alors les projections de v sur les espaces W_q.

$$\xi_q = A_q\, v = X_q\, a_q,\ q \in Q$$

où $a_q \in R^{J_q}$ désigne le facteur canonique associé.

II.4.3 Généralisation

Différentes variantes ont été présentées par des auteurs pour généraliser la méthode de Caroll. On peut citer les travaux de Kobilinski [Kob 77] qui remplace dans l'étude de Q groupes de variables, chaque groupe par un ensemble de variables orthogonales et ceux de Tennenhaus [Ten 84] qui étudie le problème d'optimisation (II.4.1) en imposant diverses contraintes linéaires aux facteurs. L'Analyse en Composantes Principales par rapport à des variables Instrumentales d'Escoufier où l'on cherche une métrique telle que l'opérateur associé aux triplets soit le plus proche d'un opérateur donné peut-être considérée comme une généralisation de l'A.C.G. de Caroll (le vecteur v est remplacé par un opérateur).

II.5 L'analyse Factorielle Multiple [EsP 82], [EsP 84]

II.5.1 Introduction

Escofier - Pages font remarquer que l'analyse canonique généralisée de Caroll pose des problèmes d'interprétation. En effet les vecteurs canoniques $\xi_q = A_q\, v$ associés à chaque groupe X_q peuvent exprimer une variance très faible de ces groupes. Ils proposent donc de chercher des combinaisons linéaires de variables d'un groupe décrivant mieux ces groupes au sens de la variance expliquée.

II.5.2 Présentation

Ils proposent de prendre comme liaison entre une variable v et un groupe X_q : l'inertie en projection du nuage des variables N_I^q sur le vecteur v. Pour cela, on

se restreint aux métriques diagonales $M_q = \{M_q^j \mid j \in J_q\}$ sur les groupes de variables. Alors la liaison L^2 d'Escofier Pages est :

$$L^2(v, X_q) = \Sigma \{M_q^j (\langle v, \underline{x}^j \rangle)^2 \mid j \in J_q\}$$

$$= I(N^q_{\Delta_v^{\perp} I})$$

inertie du nuage des variables du groupe q par rapport à l'hyperplan $\Delta_v^{\perp I}$ orthogonal à v.

Il s'agit donc pour étudier la liaison entre les Q groupes de variables de chercher une variable $v \in \mathbf{R}^I$ solution du problème d'optimisation :

Problème II.5.1

$$\left|\begin{array}{l} \max \Sigma \{L^2(v, X_q) \mid q \in Q\} \\ v \in R^I \\ \text{avec } \| v \|^2_{D_p} = 1 \end{array}\right.$$

Cette méthode revient à remplacer l'opérateur de projection A_q par la forme quadratique d'inertie $W_q D_p$ qui a même image E_q mais qui tient compte de l'inertie de la projection des variables dans les différentes directions de E_q. Le problème (II.5.1) s'écrit :

Problème II.5.1

$$\left|\begin{array}{l} \max \Sigma \{\langle v, W_q D_p v \rangle_{D_p} \mid q \in Q\} \\ v \in R^I \\ \text{sous la condition } \| v \|^2_{D_p} = 1 \end{array}\right.$$

ou encore :

Problème II.5.1

$$\left|\begin{array}{l} \max \langle v, \Sigma \{W_q D_p \mid q \in Q\} (v) \rangle_{D_p} \\ v \in R^I \\ \text{sous la condition } \| v \|^2_{D_p} = 1 \end{array}\right.$$

On a alors la proposition :

Proposition II.5.1

La variable v, la plus liée aux groupes X_q au sens de L^2, est le vecteur propre de $W D_p$ associé à la plus grande valeur propre λ où $W = \Sigma \{W_q \mid q \in Q\}$ en pondérant éventuellement les W_q par les coefficients $\{c_q \mid q \in Q\}$.

La démarche étant identique à celle de Caroll c'est-à-dire, on cherche une variable générale v la plus liée aux groupes, puis cette variable étant obtenue, les variables canoniques ξ_q s'en déduisent par application des opérateurs $W_q D_p$: $\xi_q = W_q D_p v$.

III METHODOLOGIE ET CADRES DE REFERENCE

III.1 Introduction

Après avoir examiné quelques méthodes existantes, nous présentons notre approche en introduisant les différents concepts et cadres de référence. Nouss avons déjà présenté les espaces $(F^* \otimes F, \langle\ \rangle)$ et $(E^* \otimes E, \langle\ \rangle)$ et les tenseurs U et Z associés à un triplet (X, M, N) au paragraphe II.3.5. Nous présentons l'espace $(E \otimes F, M \otimes N)$ et le tenseur $X_{E \otimes F}$ associé à un tableau X. Nous proposons et justifions la définition de mesure d'information associée à un triplet (X, M, N)

comme la norme $\| X_{E \otimes F} \|^2_{M \otimes N}$. Une telle définition a été déjà proposée par Govaert [Gov 83], pour l'étude d'un tableau de mesures par la classification croisée en se plaçant dans l'espace des applications linéaires, $L(E^*, F)$, nous avons, dans l'introduction, précisé les raisons pour lesquelles nous préférons étudier le tenseur $X_{E \otimes F}$ plutôt que l'application linéaire $X \in L(E^*, F)$. Nous montrons que cette mesure d'information s'exprime comme la trace des tenseurs U et Z et étudions les différents problèmes d'optimisation relatifs à la réduction de cette mesure d'information. Les liens entre les tenseurs $X_{E \otimes F}$, U et Z sont enfin précisés.

III.2 Mesure d'information associée un triplet

III.2.1 L'espace $(E \otimes F, M \otimes N)$

La donnée d'un tableau $X(n \times p)$ définit une application linéaire de l'espace $E^* \simeq R^{p*}$ dans $F = R^n$ de la manière suivante :

$$E^* \xrightarrow{\quad X \quad} F$$

$$e^{j*} \longrightarrow X(e^{j*}) = \underline{x}^j \text{ variable } j.$$

Ainsi, $X \in L(E^*, F)$ et $X' \in L(F^*, E)$. Rappelons que l'on a l'isomorphisme canonique $L(E^*, F) \simeq L(F^*, E) \simeq E \otimes F$ et les applications X et X' sont les expressions d'un même tenseur de $E \otimes F$ que l'on notera $X_{E \otimes F}$ qui s'exprime, relativement à la base $\{e_j \otimes f_i \mid j \in J, i \in I\}$ de $E \otimes F$:

$$X_{E \otimes F} = \Sigma \{x_i^j \, e_j \otimes f_i \mid j \in J, i \in I\}$$

L'espace E étant muni d'une métrique M et F d'une métrique N, le produit tensoriel $E \otimes F$ se trouve muni d'une structure euclidienne par la métrique produit $M \otimes N$ définit comme suit :

$$\forall (x,x') \in E, \forall (y,y') \in F \quad M \otimes N(x \otimes y, x' \otimes y') = M(x,x')N(y,y') \quad \text{(III.2.1)}$$

Donnons l'expression du produit scalaire de deux tenseurs T et S de E ⊗ G. Soient leurs expressions tensorielles :

$$T = \Sigma \{T_{ij}\ e_j \otimes f_i \mid j \in J,\ i \in I\}$$

$$S = \Sigma \{S_{i'j'}\ e_{j'} \otimes f_{i'} \mid j \in J,\ i' \in I\}$$

L'expression (III.2.1) du produit scalaire s'écrit alors :

$$< T, S >_{M \otimes N} = \Sigma \{T_{ij}\ S_{i'j'}\ M^{jj'}\ N_{ii'} \mid j,j' \in J;\ i,i' \in I\}$$

III.2.2 Mesure d'information associée à un tableau

Définition III.2.2

La mesure d'information associée à un tableau x est la norme du tenseur $X_{E \otimes F}$, le produit tensoriel E ⊗ F étant muni de la métrique M ⊗ N.

Cette mesure d'information est notée I(X, M, N) ou simplement I(X) quand aucune confusion n'est à craindre relativement au choix des métriques.

(III.2.2.1)

$$I(X) = \| X_{E \otimes F} \|^2_{M \otimes N} = \Sigma \{M^{jj'}\ N_{ii'}\ x_i^j\ x_{i'}^{j'} \mid j,j' \in J\ ;\ i,i' \in I\}$$

Commentons cette définition. Si l'on calcule la distance entre deux individus $\underline{x}_i$ et $\underline{x}_{i'}$:

$$d^2_M (\underline{x}^j, \underline{x}^{j'}) = \Sigma \{N_{ii'}\ (x_i^j - x_{i'}^{j'})^2 \mid j,j' \in J\}$$

La métrique M peut s'interpréter comme un ensemble de pondérations $\{M^{jj'} \mid j,j' \in J\}$ associées aux couples de colonnes (j, j'). De même, la distance entre deux variables $\underline{x}^j$ et $\underline{x}^{j'}$ s'écrit :

$$d^2_N(\underline{x}^j, \underline{x}^{j'}) = \Sigma \{N_{ii'} (x^j_i - x^{j'}_{i'})^2 \mid i,i' \in I\}$$

La métrique N est un ensemble de pondérations $\{N_{ii'} \mid i,i' \in I\}$ associée aux couples de lignes (i, i').

I(X) s'interpréte comme la somme des produits deux à deux des cases, (i, j) et (i', j') de X pondérés par les poids accordés aux couples (i, i') de lignes et (j, j') de colonnes.

Nous allons montrer que la quantité d'information I(X, M, N), suivant le tableau X et les métriques M, N, généralise les indicateurs statistiques usuels de variance, et inertie.

- Si le tableau se réduit à une variable : $X = \underline{x}^j$, les métriques étant M = Id et $N = D_p$ alors :

$$I(X) = I(\underline{x}^j) = \Sigma \{p_i x^{j^2}_i \mid i \in I\} = \text{var}(\underline{x}^j)$$

variance de la variable $\underline{x}^j$, si $\underline{x}^j$ est centrée.

- Pour les mêmes métriques M et N ci-desssus, si $X = (\underline{x}^j, j \in J)$, les variables étant centrées :

$$I(X) = \Sigma \{p_i x^{j2}_i \mid i \in I, j \in J\} = \Sigma \{ \text{var } \underline{x}^J \mid j \in J \}$$

En choisissant $M = D_{1/\sigma^2}$

$$I(X) = \Sigma \left\{ \frac{p_i}{\text{var } \underline{x}^j} x^{j2}_i \mid i \in I, j \in J \right\} = \text{card } J$$

- Pour un tableau général X(n × p), en choisissant $N = D_p$ et M une métrique quelconque :

$$I(X) = \Sigma \{p_i M^{jj'} x^j_i x^{j'}_i \mid i \in I ; j,j' \in J\}$$
$$= \Sigma \{p_i d^2_M (\underline{x}_i, \underline{x}_i) \mid i \in I) \mid i \in I\} = I_o(N^I_J)$$

I(X) s'interpréte comme l'inertie du nuage des individus N_J^I R^J par rapport à l'origine 0.

Si l'on choisit $M = \Delta = \begin{bmatrix} q_1 & & \\ & \ddots & \\ & & q_p \end{bmatrix}$ métrique diagonale aussi

alors :

$$I(X) = \Sigma \{p_i\, q_j\, x_i^{j2} \mid i \in I,\ j \in J\} = \Sigma \{q_j\, d^2_{D_p}(\underline{x}^j, \underline{x}^j) \mid j \in J\}$$

$= I_0(N_I^J)$ inertie du nuage des variables N_I^J R^I par rapport à l'origine 0.

Ainsi, lorsque les métriques M et N sont diagonales, I(X) représente, à la fois, l'inertie du nuage des variables N_J^I et celui des individus N_I^J :

$$\boxed{I(X) = I_0(N_I^J) = I_0(N_J^I)}$$

Nous allons donner d'autres expressions de mesure de l'information I(X) en particuliers, on trouvera que I(X) est la trace d'une application linéaire donc indépendant des bases ayant servi à le calculer.

III.2.3 Interprétation par la trace des opérateurs

Nous allons montrer que la mesure d'information I(X) s'exprime simplement en fonction des opérateurs U et Z associés au triplet (X, M, N).

Nous avons associé au triplet (X, M, N), les tenseurs $U \in F^* \otimes F$ et $Z \in E^* E$, dont les expressions (II.3.5.1) et (II.3.5.2) sont :

$$U = \Sigma \{M^{jj'}\ \underline{x}^{j*} \otimes \underline{x}^{j'} \mid j, j' \in J\}$$

et

$$Z = \Sigma \{N_{ii'}\ \underline{x}_i^* \otimes \underline{x}_{i'} \mid i, i' \in I\}$$

On note : $e = \Sigma \{e_j^* \otimes e_j \mid j \in J\}$, le tenseur unité de l'espace $E^* \otimes E$ et $f = \Sigma \{f_i^* \otimes f_i \mid i \in I\}$, celui de $F^* \otimes F$. Rappelons que e et f sont indépendantes des bases choisies $\{e_j \mid j \in J\}$ et $\{f_i \mid i \in I\}$.

On a alors la proposition :

Proposition II.2.3.1

La mesure d'information I(X) a pour expresssion :

$$I(X) = \text{trace } U = \text{Trace } Z = \langle U, f \rangle = \langle Z, e \rangle$$

Démonstration

En utilisant la définition de la trace comme contraction de produits tensoriels, on a :

$$\text{trace } U = \Sigma \{M^{jj'} \langle \underline{x}^j, \underline{x}^{j'} \rangle_N \mid j,j' \in J\}$$

$$= \Sigma \{M^{jj'} N_{ii'} x_i^j x_{i'}^{j'} \mid j,j' \in J ; i,i' \in I\}$$

$$= I(X)$$

de même :

$$\text{trace } Z = \Sigma \{N_{ii'} \langle \underline{x}_i, \underline{x}_{i'} \rangle_M \mid i,i' \in I\}$$

$$= \Sigma \{M^{jj'} N_{ii'} x_i^j x_{i'}^{j'} \mid j,j' \in J ; i,i' \in I\}$$

$$= I(X)$$

Si on considère les expressions matricielles de U et Z, on a :

$$I(X) = \text{trace } (U) = \text{trace } (X\, M\, X'\, N) = \text{trace } (X'\, N\, X\, M) = \text{trace } (Z)$$

On aurait pu trouver le résultat en utilisant la propriété que le produit scalaire entre deux tenseurs T et S dans $E \otimes F$ s'écrit :

$$< T, S >_{M \otimes N} = \text{trace } (T\ M\ S'\ N)$$

Calculons $< U, f >$

$$< U, f > = < \Sigma \{M^{jj'}\ \underline{x}^{j*} \otimes \underline{x}^{j'} \mid j,j' \in J,\ \Sigma \{f_i^* \otimes f_i \mid i \in I\} >$$

$$= \Sigma \{M^{jj'} < \underline{x}^j, \underline{f}_i >_N f_i^* (\underline{x}^{j'}) \mid i \in I\ ;\ j,j' \in J\}$$

$$= \Sigma \{M^{jj'}\ N_{ii'}\ x_{i'}^j\ x_i^{j'} \mid i,i' \in I\ ;\ j,j' \in J\}$$

$$= I(X)$$

par un calcul symétrique $I(X) = < Z, e >$

III.3 REDUCTION DE LA MESURE D'INFORMATION

III.3.1 Introduction

Les différentes expressions I(X) dans les cadres de références $(E \otimes F, M \otimes N)$, $(F^* \otimes F, < >)$ et $(E^* \otimes E, < >)$ montrent que la réduction de I(X) s'interprétera différemment suivant ces espaces. Nous adopterons deux points de vue.

III.3.2 Approximation d'un tenseur d'ordre K dans l'espace $(E \otimes F, M \otimes N)$

Tel est en effet la première formulation naturelle de l'étude de la réduction de $I(X) = \| X_{E \otimes F} \|^2_{M \otimes N}$. Il s'agit de chercher un tenseur T de rang K avec $K \leq \inf (\text{card } I, \text{card } J)$ le "plus proche possible" du tenseur $X_{E \otimes F}$ et constituant donc une bonne approximation de $X_{E \otimes F}$. Le problème d'optimisation s'écrit :

Problème III.3.2.1

$$\left| \begin{array}{l} \min \| X_{E \otimes F} - T \|^2_{M \otimes N} \\ T \in E \otimes F \\ \text{rang } T = K \end{array} \right.$$

Un tel problème a été déjà posé et résolu par Benzecri [Ben 73] dans la présentation de l'analyse factorielle comme la recherche des invariants d'un tenseur. Il montre que cela revient à étudier les invariants simultanés de deux formes quadratiques et que dans le cadre de l'analyse factorielle, par exemple, la solution est fournie par le tenseur $T = \Sigma \{\sqrt{\lambda_k}\ \phi_k \otimes \psi_k \mid k \in K\}$ où (ϕ_k, ψ_k) est le couple kème de facteurs de l'analyse des correspondances du tableau X.

La formule de reconstitution des données en est la conséquence immédiate. Sachant le problème résolu, nous nous intéressons surtout aux différentes expressions du problème d'optimisation dans les autres espaces de référence.

III.3.3 Ensembles de variables liées au sens de $\mathcal{L}$ dans $(F^* \otimes F, < >)$

Les méthodes factorielles ont été historiquement la recherche de "composantes principales" c'est-à-dire d'un ensemble de variables les plus représentatives du tableau des données X. En Analyse de Composantes Principales, ces variables sont les plus liées au sens de la corrélation à l'ensemble des variables (cf [Ca P 76]). Les composantes principales étant deux à deux non corrélées et de pouvoirs de représentativité décroissants.

En s'inspirant de ces démarches, on représente le triplet (X, M, N) par le tenseur $U \in F^* \otimes F$ et l'on cherche un ensemble de vecteurs $\{v_k \mid k \in K\}$ normés, deux à deux orthogonaux les plus liés dans un certain sens que l'on va préciser au tableau X.

On définit la liaison $\mathcal{L}$ entre un vecteur v et un triplet (X, M, N) comme le produit scalaire entre les tenseurs représentatifs des triplets (X, M, N) et (v, Id_E, N) :

$$\mathcal{L}(v, X, M) = < U, v^* \otimes V >$$

en effet, le N-projecteur $v^* \otimes v$ et le tenseur associé à (v, Id, N).

On cherchera donc un ensemble de vecteurs ordonnés $\{v_k \mid k \in K\}$, N-orthonormés dont le pouvoir de liaison va en décroissant c'est-à-dire $v_1 \in F$ est solution du problème d'optimisation.

Problème : III.3.3.1

$$\left|\begin{array}{l} \max \mathcal{L}(v_1, X, M) = \langle U, v_1^* \otimes v_1 \rangle \\ v_1 \in F \\ \| v_1 \|^2_N = 1 \end{array}\right.$$

puis $v_2 \in F$, solution de :

Problème III.3.3.2

$$\left|\begin{array}{l} \max \mathcal{L}(v_2, X, M) = \langle U, v_2^* \otimes v_2 \rangle \\ v_2 \in F \\ \| v_2 \|^2_N = 1 \\ \langle v_1, v_2 \rangle_N = 0 \end{array}\right.$$

et d'une manière générale $v_k \in F$ sera solution de :

Problème III.3.3.3

$$\left|\begin{array}{l} \max \mathcal{L}(v_k, X, M) = \langle U, v_k^* \otimes v_k \rangle \\ \| v_k \|^2_N = 1 \\ \langle v_k, v_{k'} \rangle_N = 0 \text{ pour } k' \in K \text{ et } k \neq k' \end{array}\right.$$

Soit $\{v_k \mid k \in K\}$, un ensemble de tels vecteurs liés au sens de $\mathcal{L}$ au triplet (X, M, N). Cet ensemble est aussi solution du problème suivant :

Problème III.3.3.4

$$\left|\begin{array}{l} \max \langle U, \Sigma \{v_k^* \otimes v_k \mid k \in K\} \rangle \\ v_k \in F \\ \text{tel que } \langle v_k, v_{k'} \rangle = \delta^k_{k'} \quad \forall k,k' \in K \end{array}\right.$$

En effet, supposons qu'il existe un ensemble de vecteurs $\{w_k \mid k \in K\}$, solution du problème (III.3.3.4) tel que :

$$< U, \Sigma \{w_k^* \otimes w_k \mid k \in K\} >> < U, \Sigma \{v_k^* \otimes v_k \mid k \in K\} >$$

Les vecteurs w_k étant rangés suivant la liaison $\mathcal{L}$ c'est-à-dire :

$$\mathcal{L}(w_1, X, M) \geq \mathcal{L}(w_2, X, M) \geq \ldots \geq \mathcal{L}(w_k, X, M)$$

Par hypothèse, v_1 est tel que $\mathcal{L}(v_1, X, M) \geq \mathcal{L}(w_1, X, M)$ de même v_2 est tel que $\mathcal{L}(v_2, X, M) \geq \mathcal{L}(w_2, X, M)$ car $< v_1, v_2 > = 0$ et $< w_1, w_2 > = 0$ et

$\| v_2 \|^2_N = \| w_2 \|^2_N = 1$ de même $\mathcal{L}(v_k, X, M) \geq \mathcal{L}(w_k, X, M)$ d'où :

$$\Sigma \{\mathcal{L}(v_k, X, M) \mid k \in K\} \geq \Sigma \{\mathcal{L}(w_k, X, M) \mid k \in K\}$$

ou encore :

$$\Sigma \{< U, v_k^* \otimes v_k > \mid k \in K\} \geq \Sigma \{< U, w_k^* \otimes w_k > \mid k \in K\}$$

c'est-à-dire :

$$< U, \Sigma \{v_k^* \otimes v_k \mid k \in K\} > \geq < U, \Sigma \{w_k^* \otimes w_k \mid k \in K\} >$$

ce qui est contraire à l'hypothèse de départ.

Ainsi, un ensemble de vecteurs $\{v_k \mid k \in K\}$ liés au sens de $\mathcal{L}$ au triplet (X, M, N) est solution du problème (III.3.3.4), la réciproque n'étant pas vraie. En effet toute base N-orthornormée $\{W_k \mid k \in K\}$ du sous-espace $F_k = \Sigma\{v_k^* \otimes v_k \mid k \in K\}$ (F) est solution du problème (III.3.3.4) mais n'est pas liée au sens de $\mathcal{L}$ à (X,M,N).

Nous allons montrer l'équivalence des problèmes (III.3.2.1) et (III.3.3.4), en établissant le lemme suivant :

Lemme : III.3.3.1

Soit $F_K \subset F$, un sous-espace vectoriel de dimension K, $\{v_k \mid k \in K\}$, une base N-orthonormée de F_K. On note $\tilde{X}_{E \otimes F_K}$, la projection $M \otimes N$ orthogonale de $X_{E \otimes F}$ sur $E \otimes F_K$.

On a alors le relation :

$$\| \tilde{X}_{E \otimes F_K} \|^2_{M \otimes N} = < U, \Sigma \{v_k^* \otimes v_k \mid k \in K\} > \qquad \text{(III.3.3.1)}$$

On note $A^{E \otimes F}_{E \otimes F_k}$, le projecteur N-orthogonal de $E \otimes F$ sur $E \otimes F_K$:

$$\tilde{X}_{E \otimes F_K} = A^{E \otimes F}_{E \otimes F_K} (X_{E \otimes F})$$

on sait que : $A^{E \otimes F}_{E \otimes F_K} = A^E_E \otimes A^F_{F_K} = Id_E \otimes A^F_{F_K}$ où $A^F_{F_K}$ est le projecteur N-orthogonal de F sur F_K, d'après les propriétés du produit tensoriel d'applications linéaires. Soit $\tilde{X}$, l'application linéaire de E^* dans F_K associée au tenseur $\tilde{X}_{E \otimes F_K}$, $\tilde{X} \in L(E^*, F_K)$, X celle associée au tenseur $X_{E \otimes F}$, $X \in L(E^*, F)$, on a alors la relation :

$$\tilde{X} = A^F_{F_K} X$$

Désignons par $A(k \times n)$, la matrice relative à $A^F_{F_K}$. Les tenseurs $\tilde{X}_{E \otimes F_K}$ et $X_{E \otimes F}$ étant $M \otimes N$ orthogonaux dans $E \otimes F$, on a :

$$< \tilde{X}_{E \otimes F_K}, \tilde{X}_{E \otimes F_K} >_{M \otimes N} = < \tilde{X}_{E \otimes F_K}, X_{E \otimes F} >_{M \otimes N}$$

or on a vu que :

$$< \tilde{X}_{E \otimes F_K}, X_{E \otimes F} >_{M \otimes N} = \text{trace } (A \; X \; M \; X' \; N)$$

On sait que le projecteur $A^F_{F_K} \in L(F,F)$ peut s'exprimer tensoriellement en fonction de la base N-orthonormée $\{v_k \mid k \in K\}$ de F_K comme suit :

$$A^{F}_{F_K} = \Sigma \{ v_k^* \otimes v_k \mid k \in K \}$$

ou matriciellement : $A = \Sigma \{ v_k \; v'_k \; N \mid k \in K \}$ d'où :

$$< \tilde{X}_{E \otimes F_k}, X_{E \otimes F} >_{M \otimes N} = \text{trace} \left(\Sigma \{ v_k \; v'_k \; N \mid k \in K \} \; X \; M \; X' \; N \right)$$

$$= < U, \Sigma \{ v_k^* \otimes v_k \mid k \in K \} >$$

d'où le lemme :

$$\| \tilde{X}_{E \otimes F_K} \|^2_{M \otimes N} = < U, \Sigma \{ v_k^* \otimes v_k \mid k \in K \} >$$

On a alors la proposition relative à l'équivalence des problèmes (III.3.2.1) et (III.3.3.4) :

Proposition : III.3.3.1

Si $\{ v_k \mid k \in K \}$ est un ensemble de vecteurs de F, N-orthonormés solutions du problème d'optimisation (III.3.3.4), alors si on désigne par F_K le sous-espace de dimension K engendré par les vecteurs v_k alors $\tilde{X}_{E \otimes F_K}$, projection M ⊗ N orthogonale de $X_{E \otimes F}$ sur $E \otimes F_K$ est la meilleure approximation de rang K du tenseur $X_{E \otimes F}$ donc solution du problème (III.3.2.1).

Réciproquement : si $\tilde{X}_{E \otimes F}$ est la meilleure approximation de rang K du tenseur $X_{E \otimes F}$, alors si on note $F_K = \tilde{X}(E^*)$, le sous-espace vectoriel de dimension K image de E^* par $\tilde{X}$, soit l'ensemble $\{ v_k \mid k \in K \}$, une base N-orthonormée de F_K alors l'ensemble $\{ v_k \mid k \in K \}$ est solution du problème (III.3.3.4).

Démonstration : Les problèmes d'optimisation (III.3.2.1) et (III.3.3.4) sont équivalents. En effet, si l'ensemble $\{ v_k \mid k \in K \}$ est une solution du problème (III.3.2.1), on ne peut trouver un tenseur $Y_{E \otimes F}$ meilleure approximation de rang K de $X_{E \otimes F}$ tel que :

$$\| Y_{E \otimes F} \|^2_{M \otimes N} > \| \tilde{X}_{E \otimes F} \|^2_{M \otimes N}$$

En effet, en considérant l'espace G_K de dimension K image de Y de E^*, c'est-à-dire : $G_K = Y(E^*)$ nécessairement $Y_{E \otimes F} = A_{E \otimes G_K}^{E \otimes F}(X_{E \otimes F}) = X_{E \otimes G_K}$ c'est-à-dire la projection $M \otimes N$ orthogonale de $X_{E \otimes F}$ sur $E \otimes G_K$ (sinon ce tenseur projection serait une meilleure approximation de rang K que $Y_{E \otimes F}$) par suite en vertu du lemme (III.3.3.1), on trouverait un ensemble de vecteurs $\{w_k \mid k \in K\}$, N-orthonormés base de G_K tel que :

$$\| X_{E \otimes G_K} \|^2_{M \otimes N} = < U, \Sigma \{ w_k \overset{*}{\otimes} w_k \mid k \in K \} >> < u, \Sigma \{ v_k \overset{*}{\otimes} v_k \mid k \in K \} >$$

ce qui est contraire à l'hypothèse.

<u>Réciproquement</u> : si $X_{E \otimes F_K}$ est la meilleure approximation de rang K de $X_{E \otimes F}$, $\{v_k \mid k \in K\}$ étant toujours une base N-orthonormés de F_K, cet ensemble est solution du problème (III.3.3.4). On ne peut trouver un ensemble $\{w_k \mid k \in K\}$ de vecteurs N-orthonormés tel que :

$$< U, \Sigma \{ w_k \overset{*}{\otimes} w_k \mid k \in K \} > > < U, \Sigma \{ v_k \overset{*}{\otimes} v_k \mid k \in K \} >$$

Sinon en considérant G_K, l'espace engendré par les $\{w_k \mid k \in K\}$, on trouverait un tenseur de rang K :

$$X_{E \otimes G_K} = A_{E \otimes G_K}^{E \otimes F} (X_{E \otimes F})$$

tel que :

$$\| X_{E \otimes G_K} \|^2_{M \otimes N} = <U, \Sigma \{ w_k \overset{*}{\otimes} w_k \mid k \in K \}> > \| X_{E \otimes F_K} \|^2_{M \otimes N} = <U, \Sigma \{ v_k \overset{*}{\otimes} v_k \mid k \in K \}>$$

ce qui est contraire à l'hypothèse.

La résolution de ces problèmes d'optimisations classiques dans les méthodes factorielles revient à l'étude des invariants simultanés de deux formes quadratiques (cf [Ben 73]) et montreront que le sous-espace F_K est unique (à condition que les variables propres ne soient pas multiples) engendré par les vecteurs propres $\{v_K \mid k \in K\}$ de U formant une base N-orthonormée de F.

Les vecteurs $\{v_k \mid k \in K\}$ ne seront pas seulement une solution du problème (III.3.2.1) mais seront un ensemble de vecteurs les plus liés au sens de $\mathfrak{f}$ au triplet (X, M, N) (solutions de III.3.3.1, III.3.3.2, III.3.3.3).

Nous avons raisonné dans l'espace $F^* \otimes F$, de manière symétrique, on a la relation en s'intéressant à l'espace $E^* \otimes E$:

$$\| \tilde{X}_{E_K \otimes F} \|^2_{M \otimes N} = < Z, \Sigma \{u_k^* \otimes u_k \mid k \in K\} >$$

où $\{u_k \mid k \in K\}$ est une base M-orthonormée de l'espace E_K, image de l'application X' de F^* : $E_K = X'(F^*)$. On sait que les opérateurs U et Z ont les mêmes valeurs propres par suite, la meilleure approximation de rang K du tenser $X_{E \otimes F}$ est la projection de ce tenseur sur $E_K \otimes F_K$ où E_K, l'espace vectoriel engendré parles K premiers vecteurs propres de Z et F_K, celui engendré par les K premiers vecteurs propres de U. On a donc :

$$\boxed{\| \tilde{X}_{E_K \otimes F} \|^2_{M \otimes N} = <Z, \Sigma\{u_k^* \otimes u_k \mid k \in K\} > = <U, \Sigma\{v_k^* \otimes v_k \mid k \in K\}>}$$

comme : $Z = \Sigma\{\lambda_k \, u_k^* \otimes u_k \mid k \in K\}$, $<Z, \Sigma\{u_k^* \otimes u_k \mid k \in K\}> = \Sigma\{\lambda_k \mid k \in K\}$

il vient :

$$\boxed{\| \tilde{X}_{E_K \otimes F_K} \|^2_{M \otimes N} = \Sigma\{\lambda_k \mid k \in K\}}$$

Nous allons donner l'expression de la décomposition de la mesure de l'information I(X) dans les espaces $E^* \otimes E$ et $F^* \otimes F$.

Proposition : III.3.3.2

Pour tout vecteur $v \in \mathbb{R}^I$ et $u \in \mathbb{R}^J$ de normes unitaires $\| u \|^2_M = \| v \|^2_N = 1$, on a la décomposition dans les espaces $F^* \otimes F$ et $E^* \otimes E$

$$I(X) = \langle U, v^* \otimes v \rangle + \langle u, \otimes^{\perp_N} v \rangle$$

et

$$I(X) = \langle Z, u^* \otimes u \rangle + \langle Z, \otimes^{\perp_M} u \rangle$$

où :

$$\otimes^{\perp_N} v = \Sigma \{ b_i^* \otimes b_i \mid i \in [1, \ldots, I-1] \}$$

et

$$\otimes^{\perp_M} u = \Sigma \{ a_j^* \otimes a_j \mid j \in [1, \ldots, J-1] \},$$

les b_i (respectivement a_j) sont les vecteurs N-orthonormés (respectivement M-orthor- normés) de l'hyperplan $\Delta v^{\perp_N}$ (respectivement $\Delta u^{\perp_M}$) N-orthogonal (respecti- vement M-orthogonal) à Δv (respectivement Δu) droite engendrée par v (respectivement u).

Démonstration :

Nous avons vu que $I(X) = \langle U, f \rangle$, il suffit de montrer que le tenseur unité f de $F^* \otimes F$ admet la décomposition

$$f = v^* \otimes v + \otimes^{\perp_N} v$$

égalité conséquence de la décomposition N-orthonormale de l'espace F, en effet,

$$F = \underset{N}{\Delta v \otimes \Delta v^{\perp_N}}$$

par suite, $\forall\, y \in F$, on a :

$$\underline{y} = \langle \underline{y}, v \rangle_N v + \Sigma \{ \langle \underline{y}, b_i \rangle_N b_i \mid i \in [1, \ldots, I-1] \}$$

en effet, $\{v, b_i \mid i \in [1, \ldots, I-1]\}$ forment une base N-orthonormale de F d'où le résultat :

$$f = v^* \otimes v + \Sigma \{ b_i^* \otimes b_i \mid i \in [1, \ldots, I-1] \}$$

la linéarité du produit scalaire permet d'avoir le résultat. La démonstration est symétrique pour la deuxième expression.

L'expression géométrique de ces deux expressions dans $E \otimes F$ est donnée par la proposition :

Proposition : III.3.3.3

on a dans $E \otimes F$:

$$\| X_{E \otimes F} \|^2_{M \otimes N} = \| X_{E \otimes \Delta v} \|^2_{M \otimes N} + \| X_{E \otimes \Delta v^{\perp N}} \|^2_{M \otimes N}$$

$$= \| X_{\Delta u \otimes F} \|^2_{M \otimes N} + \| X_{\Delta u^{\perp M} \otimes F} \|^2_{M \otimes N}$$

Ces expressions ne sont que la conséquence du théorème de Pythagore appliqué aux sous-espaces $E \otimes \Delta v$ et $E \otimes \Delta v^{\perp N}$ somme directe de $E \otimes F$, $M \otimes N$ orthogonaux. En vertu du lemme, on a montré en effet que :

$$\| X_{E \otimes \Delta v} \|^2_{M \otimes N} = < U, v^* \otimes v >$$

et

$$\| X_{\Delta u \otimes F} \|^2_{M \otimes N} = < Z, u^* \otimes u >$$

on a alors :

$$\| X_{E \otimes \Delta v^{\perp N}} \|^2_{M \otimes N} = < U, \otimes^{\perp N} v >$$

et

$$\| X_{\Delta u^{\perp M} \otimes F} \|^2_{M \otimes N} = < Z, \otimes^{\perp M} v >$$

Au terme de cette présentation, nous avons introduit la définition de la mesure d'information associée à un triplet (X, M, N), montré quelques aspects de la réduction de cette information en termes géométriques (approximation d'ordre K d'un tenseur) et de la recherche de variables liées à un triplet au sens de $\mathcal{L}$, avant d'étudier la généralisation à un ensemble de triplets, nous allons préciser les relations entre les différents tenseurs $X_{E \otimes F}$, U et Z relatifs à (X, M, N).

III.3.4 Liens entre les tenseurs $X_{E \otimes F}$, U et Z relatifs à (X, M, N)

A un tableau X, nous avons associé différents tenseurs et espaces de références, ce sont :

le tenseur $X_{E \otimes F}$ de l'espace $(E \otimes F, M \otimes N)$

le tenseur U de l'espace $(F^* \otimes F, < >)$

et le tenseur Z de l'espace $(E^* \otimes E, < >)$

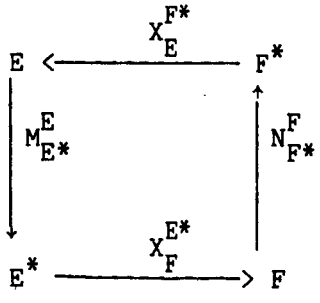
Nous précisons dans cette partie, les liens existant entre $X_{E \otimes F}$, U et Z. Les expressions de ces tenseurs sont, rappelons-le :

$$X_{E \otimes F} = \Sigma \{x_i^j \, e_j \otimes f_i \mid j \in J, \; i \in I\}$$

$$U = \Sigma \{M^{jj'} \, \underline{x}^{j*} \otimes \underline{x}^{j'} \mid j, j' \in J\}$$

$$Z = \Sigma \{N_{ii'} \, \underline{x}_i^* \otimes \underline{x}_{i'} \mid i, i' \in I\}$$

Nous indiquons l'espace d'origine et l'espace but pour une application donnée, en écrivant en indice supérieur, l'espace d'origine et en indice inférieur, l'espace but. Ainsi, le schéma de dualité relatif à un tableau X s'écrit :



X_F^{E*}, X_E^{F*} étant les deux applications linéaires transposées l'une de l'autre relatives au tenseur $X_{E \otimes F}$. M_{E*}^E et N_{F*}^F étant les différentes applications linéaires induites par la donnée des métriques M et N sur les espaces E et F.

Ces différentes applications linéaires induisent des applications linéaires sur les espaces produits tensoriels formés à partir des espaces E, F et leurs duaux E^*, F^* s'organisant suivant le schéma (III.3.4.1).

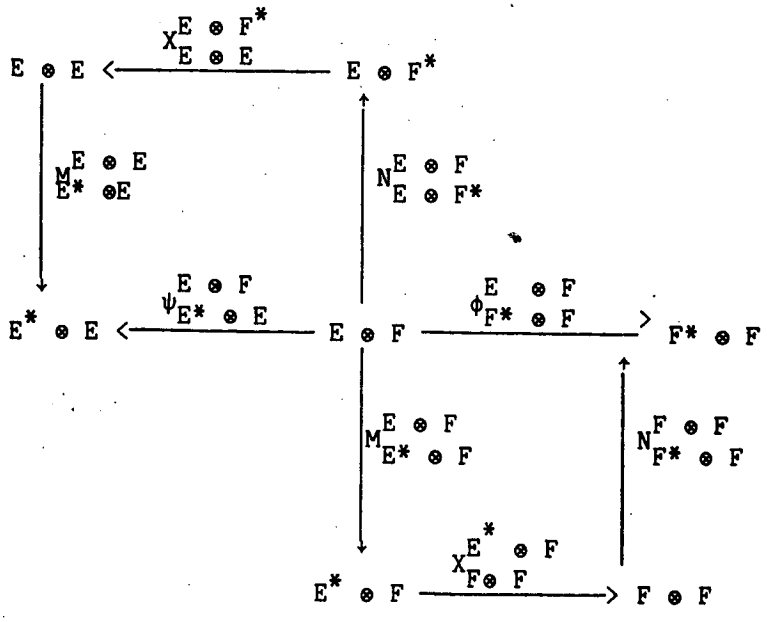


schéma III.3.4.1

Considérons l'ensemble d'applications linéaires :

$$A_X = \{X^{E^*}_F,\ X^{F^*}_E,\ M^E_{E^*},\ N^F_{F^*}\}$$

Les application entrant dans le schéma (III.3.4.1) sont les éléments du produit tensoriel :

$$A_X \otimes Id^F_F = \{f \otimes Id^F_F \mid f \in A_X\}$$

et

$$Id^E_E \otimes A_X = [id^E_E \otimes f \mid f \in A_X\}$$

en effet, nous avons la propriété du produit tensoriel d'applications linéaires :

$$(f \otimes g)^{E \otimes F}_{G \otimes H} = f^E_G \otimes g^F_H$$

quels que soient les espaces E, F, G, H et les applications f_G^E et g_H^F linéaires.

On note $\phi_{F^* \otimes F}^{E \otimes F} = N_{F^* \otimes F}^{F \otimes F} \circ X_{F \otimes F}^{E^* \otimes F} \circ M_{E^* \otimes F}^{E \otimes F} \in L(E \otimes F, F^* \otimes F)$

et :

$$\psi_{E^* \otimes E}^{E \otimes F} = M_{E^* \otimes E}^{E \otimes E} \circ X_{E \otimes E}^{E \otimes F^*} \circ N_{E \otimes F^*}^{E \otimes F} \in L(E \otimes F, E^* \otimes E)$$

Ces applications linéaires sont définies par la donnée d'un triplet (X, M, N) on a alors la proposition :

Proposition : III.3.4.1

Les rellations entre les tenseurs $X_{E \otimes F} \in E \otimes F$, $U \in F^* \otimes F$, et $Z \in E^* \otimes E$ associés à un tableau X sont :

$$U = \phi_{F^* \otimes F}^{E \otimes F} (X_{E \otimes F})$$

$$Z = \psi_{E^* \otimes E}^{E \otimes F} (X_{E \otimes F})$$

Démonstration

On a $X_{E \otimes F} = \Sigma\{x_i^j \, e_j \otimes f_i \mid j \in J, \, i \in I\} = \Sigma\{e_j \otimes (\Sigma\{x_i^j \, f_i \mid i \in I\}) \mid j \in J\}$

$$= \Sigma \{e_j \otimes \underline{x}^j \mid j \in J\}$$

par suite :

$$M^{E \otimes F}_{E^* \otimes F}\left(\Sigma \{e_j \otimes \underline{x}^j \mid j \in J\}\right) = \Sigma \{M^E_{E^*}(e_j) \otimes Id^F_F(\underline{x}^j) \mid j \in J\}$$

remarquons que : $M^E_{E^*}(e_j) = \langle e_j, . \rangle_M = \Sigma \{M^{jj'} e^*_{j'} \mid j' \in J\} \in E^*$ d'où :

$$M^{E \otimes F}_{E^* \otimes F}(X_{E \otimes F}) = \Sigma \{M^{jj'} e^*_{j'} \otimes \underline{x}^j \mid j,j' \in J\}$$

puis

$$X^{E^* \otimes F}_{F \otimes F} \circ M^{E \otimes F}_{E^* \otimes F}(X_{E \otimes F}) = \Sigma \{M^{jj'} X^{E^*}_F(e^*_{j'}) \otimes Id^F_F(\underline{x}^j) \mid j,j' \in J\}$$

comme $\underline{x}^{j'} = X^{E^*}_F(e^*_{j'})$ d'où :

$$X^{E^* \otimes F}_{F \otimes F} \circ M^{E \otimes F}_{E^* \otimes F}(X_{E \otimes F}) = \Sigma \{M^{jj'} \underline{x}^{j'} \otimes \underline{x}^j \mid j,j' \in J\}$$

d'où

$$N^{F \otimes F}_{F^* \otimes F} \circ X^{E^* \otimes F}_{F \otimes F} \circ M^{E \otimes F}_{E^* \otimes F}(X_{E \otimes F}) = \Sigma \{M^{jj'} N^F_{F^*}(\underline{x}^{j'}) \otimes \underline{x}^j \mid j,j' \in J\}$$

$$= \Sigma \{M^{jj'} \underline{x}^{j'*} \otimes \underline{x}^j \mid j,j' \in J\}$$

en effet, $N^F_{F^*}(\underline{x}^{j'}) = \langle \underline{x}^{j'}, . \rangle_N = \underline{x}^{j'*}$, on reconnait alors l'expression de U d'où :

$$U = N^{F \otimes F}_{F^* \otimes F} \circ X^{E^* \otimes F}_{F \otimes F} \circ M^{E \otimes F}_{E^* \otimes F}(X_{E \otimes F}) \quad \text{CQFD}$$

Par un calcul symétrique, on montre l'égalité relatif à Z.

Les applications $\phi^{E \otimes F}_{F^* \otimes F}$ et $\psi^{E \otimes F}_{E^* \otimes E}$ ne sont des isomorphismes que si $X^{E^*}_F$ ou $X^{F^*}_E$ sont injectives, d'autre part si U_X et U_Y sont deux opérateurs associés à $X_{E \otimes F}$ et $Y_{E \otimes F}$, on a :

$$\langle U_x, U_y \rangle \neq \langle X_{E \otimes F}, Y_{E \otimes F} \rangle_{M \otimes N}$$

en effet, $< X_{E \otimes F}, Y_{E \otimes F} >_{M \otimes N} = \text{trace } (X\, M\, Y'\, N)$

$$\neq \text{trace } (X\, M\, X'\, N\, Y\, MY'\, N) = \langle U_X, U_Y \rangle$$

les applications $\phi^{E \otimes F}_{F^* \otimes F}$ et $\psi^{E \otimes F}_{E^* \otimes E}$ ne conservent donc pas le produit scalaire M $\otimes$ N.

III.4 ETUDE GENERALE D'UN ENSEMBLE DE TRIPLETS

III.4.1 Introduction

Nous avons donc défini la mesure d'information relative à un triplet et étudié les différentes expressions et interprétation de la réduction de cette information dans les cadres de référence choisis.

Dans cette partie, nous abordons l'étude d'un ensemble de triplets $(X_{\alpha\beta}, M_\alpha, N_\beta)$ indicés par deux ensembles finis A et B, $\alpha \in A$ et $\beta \in B$. Le but principal de ce chapitre est de montrer que l'on peut construire mathématiquement un triplet unique (X, M, N) constitué d'un tenseur relatif à un tableau X et deux métriques M et N tel que la mesure d'information I(X) soit la somme pondérée des informations $I(X_{\alpha\beta})$ relatifs aux triplets $(X_{\alpha\beta}, M_\alpha, N_\beta)$. Les liens existant entre les tenseurs U et Z et $U_{\alpha\beta}$ et $Z_{\alpha\beta}$ sont ensuite étudiés et les différentes possibilités de réduction de l'information condidérées. Enfin, le chapitre se termine sur l'étude des tableaux croisés (exemple : tableau de contingence, Burt) qui tiennent en Analyse de Données une grande importance.

III.4.2 Décomposition de produits tensoriels d'espaces et métriques

Nous supposons que nous avons un ensemble de couples d'espaces vectoriels et métriques (E_α, M_α), $\alpha \in A$

$$\text{et } (F_\beta, N_\beta),\ \beta \in B.$$

On définit les espaces E et F sommes directes orthogonales des espaces E_α, $\alpha \in A$ et F_β, $\beta \in B$.

$$E = \Sigma \{E_\alpha \mid \alpha \in A\}$$

$$F = \Sigma \{F_\beta \mid \beta \in B\}$$

Si $\{c_\alpha \mid \alpha \in A\}$ et $\{c'_\beta \mid \beta \in B\}$ sont des pondérations associées aux ensembles (E_α, M_α) et (F_β, N_β), on définit les métriques pondérées M et N sur E et F comme suit :

$$M = \Sigma \{c_\alpha M_\alpha \mid \alpha \in A\}$$

$$N = \Sigma \{c'_\beta N_\beta \mid \beta \in B\}$$

Si l'on note les projections canoniques relatives à E et F :

$$\pi_\alpha : E \longrightarrow E_\alpha,\ \alpha \in A \quad \text{et} \quad \pi_\beta : F \longrightarrow F_\beta,\ \beta \in B$$
$$x \longrightarrow \pi_\alpha(\underline{x}) \qquad \underline{y} \longrightarrow \pi_\beta(\underline{y})$$

Les projections π_α et π_β définissent la projection $\pi_{\alpha\beta} = \pi_\alpha \otimes \pi_\beta$ de $E \otimes F$ sur $E_\alpha \otimes F_\beta$, tel que :

$$\pi_{\alpha\beta} : E \otimes F \longrightarrow E_\alpha \otimes F_\beta$$
$$\underline{x} \otimes \underline{y} \longrightarrow \pi_{\alpha\beta}(\underline{x} \otimes \underline{y}) = \pi_\alpha(x) \otimes \pi_\beta(\underline{y})$$

Rappelons que l'on a les propriétés suivants relativement au produit tensoriel $E \otimes F$. (cf[Sch 81]).

1) $E \otimes F$ est la somme directe orthogonale des espaces $E_\alpha \otimes F_\beta$

$$E \otimes F = \Sigma \{E_\alpha \otimes F_\beta \mid \alpha \in A,\ \beta \in B\}$$

2) $M \otimes N$ est une métrique pondérée des $M_\alpha \otimes N_\beta$ dans $E \otimes F$

$$M \otimes N = \Sigma \{c_\alpha\, c'_\beta\, M_\alpha \otimes N_\beta \mid \alpha \in A,\ \beta \in B\}$$

vérifions cette dernière affirmation :

$\forall\, \underline{x},\underline{x}' \in E$ et $\underline{y},\underline{y}' \in F$, on a par définition :

$$M \otimes N\,(\underline{x} \otimes \underline{y},\ \underline{x}' \otimes \underline{y}') = M(\underline{x},\underline{x}')\,N(\underline{y},\underline{y}')$$

$$= [\Sigma\{c_\alpha M_\alpha\,(\pi_\alpha(\underline{x}),\ \pi_\alpha(\underline{x}))\mid \alpha \in A]\ [\Sigma\{c'_\beta N_\beta(\pi_\beta(\underline{y}),\ \pi_\beta(\underline{y}'))\mid \beta \in B\}$$

$$= \Sigma\{c_\alpha c'_\beta M_\alpha\,(\pi_\alpha(\underline{x}),\ \pi_\alpha(\underline{x}'))\,N_\beta\,(\pi_\beta(\underline{y}),\ \pi_\beta(\underline{y}'))\mid \alpha \in A,\ \beta \in B\}$$

$$= \Sigma\{c_\alpha c'_\beta M_\alpha \otimes N_\beta\,(\pi_\alpha(\underline{x}) \otimes \pi_\beta(\underline{y}),\ \pi_\alpha(\underline{x}') \otimes \pi_\beta(\underline{y}'))\mid \alpha \in A,\ \beta \in B\}$$

$$= \Sigma\{c_\alpha c'_\beta M_\alpha \otimes N_\beta(\pi_{\alpha\beta}(\underline{x} \otimes \underline{y}),\ \pi_{\alpha\beta}(\underline{x}' \otimes \underline{y}'))\mid \alpha \in A,\ \beta \in B\}$$

ce qui achève la vérification. On a donc le lemme suivant, pour tout tenseur $X_{E\otimes F} = \pi\,\{X_{E_\alpha} \otimes E_\beta \mid \alpha \in A,\ \beta \in B\}$

lemme : III.4.2.1

On a l'égalité suivante :
(III.4.2.1)

$$\| X_{E \otimes F}\|^2_{M \otimes N} = \Sigma\,\{c_\alpha c'_\beta \,\| X_{E_\alpha} \otimes F_\beta\|^2_{M_\alpha} \otimes N_\beta \mid \alpha \in A,\ \beta \in B\}$$

III.4.3 Mesure d'information et tenseurs relatifs à un ensemble de triplets

Nous considérons ici un ensemble de triplets $(X_{\alpha\beta}, M_\alpha, N_\beta)$ (cf tableau III.4.3.1) auquel est associé les applications linéaires $X_{\alpha\beta} \in L(E^*_\alpha, F_\beta)$, $X'_{\alpha\beta} \in L(F^*_\beta, E_\alpha)$ et les tenseurs $X_{E_\alpha} \otimes F_\beta$ de $E_\alpha \otimes F_\beta$, $U_{\alpha\beta} \in F^*_\beta \otimes F_\beta$ et $Z_{\alpha\beta} \in E^*_\alpha \otimes E_\alpha$ et ceci pour $\alpha \in A$ et $\beta \in B$.

Les espaces E_α et F_β sont de dimension finis et on note $J_\alpha = \dim E_\alpha$ et $I_\beta = \dim F_\beta$ pour $\alpha \in A$, $\beta \in B$.

Soit $X = \pi \{X_{\alpha\beta} \mid \alpha \in A, \beta \in B\}$, le tableau relatif au tenseur : $X_{E \otimes F} = \pi \{X_{E_\alpha \otimes F_\beta} \mid \alpha \in A, \beta \in B\}$ qui peut être considéré comme la juxtaposition des tableaux $X_{\alpha\beta}$. Comme $E = \Sigma \{E_\alpha \mid \alpha \in A\}$, on a :

$\dim E = \Sigma \{\dim E_\alpha \mid \alpha \in A\}$ et on note $J = \Sigma \{J_\alpha \mid \alpha \in A\}$. La dimension de E qui est le nombre de colonnes du tableau X et I la dimension de F est tel que : $I = \Sigma \{I_\beta \mid \beta \in B\}$, le nombre de lignes de X. Si $\underline{x}^j = X(e_j^*) \in F$, le vecteur colonne de X, on note : $\underline{x}_\beta^j = \pi_\beta(\underline{x}^j)$, la projection de $\underline{x}^j$ sur F_β de même pour $\underline{x}_i = X'(f_i^*)$: $\underline{x}_i^\alpha = \pi_\alpha(\underline{x}_i)$, la projection de $\underline{x}_i$ sur E_α, on a alors :

$$X'_{\alpha\beta} = (\underline{x}_i^\alpha \mid i \in I_\beta)$$

et

$$X_{\alpha\beta} = (\underline{x}_\beta^j \mid j \in J_\alpha)$$

Les tenseurs associés au triplet $(X_{\alpha\beta}, M_\alpha, N_\beta)$ ont pour expression :

$$U_{\alpha\beta} = \Sigma \{M_\alpha^{jj'} \; \underline{x}_\beta^{j*} \otimes \underline{x}_\beta^{j'} \mid j,j' \in J_\alpha\}$$

$$Z_{\alpha\beta} = \Sigma \{N_\beta^{ii'} \; \underline{x}_i^{\alpha*} \otimes \underline{x}_{i'}^{\alpha} \mid i;i' \in I_\beta\}$$

D'après (II.2.1),

$$M = \{M^{jj'} = \delta_{a(j')}^{a(j)} \; c_{a(j)} \; M_{a(j)}^{jj'} \mid j,j' \in J\}$$

et

$$N = \{N^{ii'} = \delta_{b(i')}^{b(i)} \; c'_{b(i)} \; N_{b(i)}^{ii'} \mid i,i' \in I\}$$

où a, b sont des fonctions sur J et I tel que $a(j) = \alpha$ si $j \in J_\alpha$ et $b(i) = \beta$ si $i \in I_\beta$. Définissons les tableaux marginaux suivants :

$X'_{\alpha.} = \pi \{X'_{\alpha\beta} \mid \beta \in B\}$ qui correspond à un ensemble de J_α colonnes du tableau X auxquels sont associés les tenseurs $U_\alpha \in F^* \otimes F$ et $Z_\alpha \in E_\alpha^* \otimes E_\alpha$ correspondant au triplet $(X'_{\alpha.}, M_\alpha, N)$. (cf : tableau III.4.3.3)

De même, $X_{.\beta} = \pi \{X_{\alpha\beta} \mid \alpha \in A\}$, un ensemble I_β lignes du tableau X auxquels sont associés les tenseurs $U_\beta \in F_\beta^* \otimes F_\beta$ et $Z_\beta \in E^* \otimes E$ correspondant au triplet $(X_{.\beta}, M, N_\beta)$. (cf : tableau III.4.3.2)

Proposition : III.4.3.1

Les différents tenseurs vérifient les égalités suivantes :

(III.4.3.1) $$U_\beta = \Sigma \{c_\alpha U_{\alpha\beta} \mid \alpha \in A\}$$

(III.4.3.2) $$Z_\alpha = \Sigma \{c'_\beta Z_{\alpha\beta} \mid \beta \in B\}$$

Si $\pi_\beta^* \otimes \pi_\beta$ est le projecteur de $F^* \otimes F$ sur $F_\beta^* \otimes F_\beta$ et $\pi_\alpha^* \otimes \pi_\alpha$ celui relatif à $E^* \otimes E$ sur $E_\alpha^* \otimes E_\alpha$

$$\pi_\beta^* \otimes \pi_\beta (U_\alpha) = c_\alpha U_{\alpha\beta}$$

$$\pi_\alpha^* \otimes \pi_\alpha (Z_\beta) = c'_\beta Z_{\alpha\beta}$$

Si on note U, Z les tenseurs associés au triplet (X, M, N), on a :

$$U = \Sigma \{c_\alpha U_\alpha \mid \alpha \in A\}$$

$$Z = \Sigma \{c'_\beta Z_\beta \mid \beta \in B\}$$

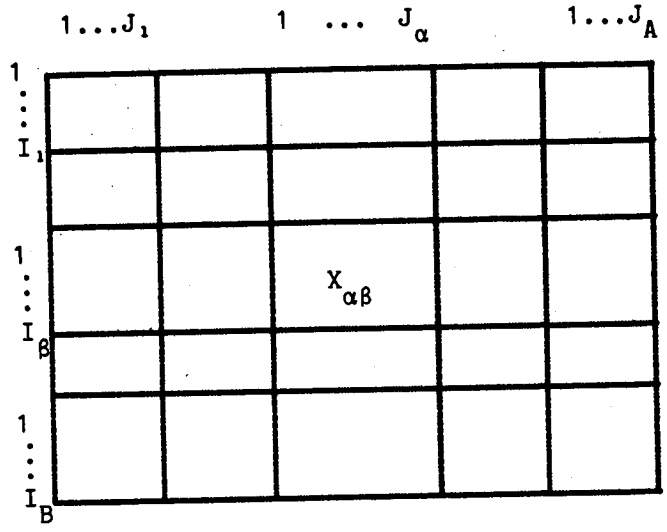


tableau général (III.4.3.1)

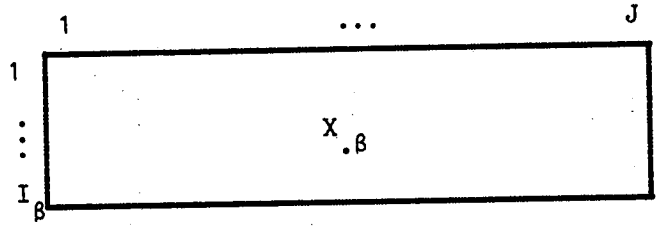


tableau histoire des variables (III.4.3.2)

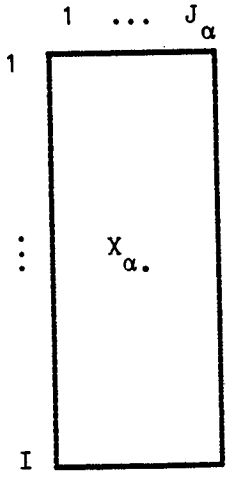
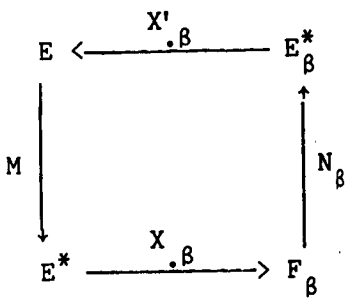
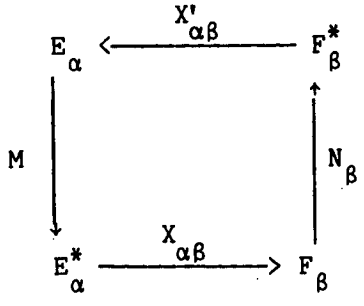


tableau histoire des individus (III.4.3.3)

Demonstration

Soit le triplet $(X_{.\beta}, M, N_\beta)$, considérons les décompositions $E = \sum\{E_\alpha \mid \alpha \in A\}$ et $M = \Sigma\{c_\alpha M_\alpha \mid \alpha \in A\}$. Les tenseurs U_β et $U_{\alpha\beta}$ appartiennent au même espace $F_\beta^* \otimes F_\beta$, les schémas de dualité associés aux triplets $(X_{.\beta}, M, N_\beta)$ et $(X_{\alpha\beta}, M_\alpha, N_\beta)$ étant :





Par définition $U_\beta = \Sigma\{M^{jj'}\ \underline{x}_\beta^{J*} \otimes \underline{x}_\beta^{j'} \mid j,j' \in J\}$, notons $\alpha = a(j)$, on a alors :

$$U_\beta = \Sigma\{\delta^{\alpha}_{a(j')}\ c_\alpha\ M_\alpha^{jj'}\ \underline{x}_\beta^{j*} \otimes \underline{x}_\beta^{j'} \mid j,j' \in J,\ \alpha \in A\}$$

$$= \Sigma\{c_\alpha\ M_\alpha^{jj'}\ \underline{x}_\beta^{j*} \otimes \underline{x}_\beta^{j'} \mid j,j' \in J_\alpha\ ;\ \alpha \in A\}$$

$$= \Sigma\{c_\alpha\ U_{\alpha\beta} \mid \alpha \in A\}$$

symétriquement, on a $Z_\alpha = \Sigma\{c'_\beta\ Z_{\alpha\beta} \mid \beta \in B\}$. Le tenseur $U_\alpha \in F^* \otimes F$, on a la décomposition $F^* \otimes F = \sum\{F_\beta^* \otimes F_{\beta'} \mid \beta, \beta' \in B\}$, calculons la projection de U_α sur l'espace $F_\beta^* \otimes F_\beta$:

$$\pi_\beta^* \otimes \pi_\beta\ (U_\alpha) = \pi_\beta^* \otimes \pi_\beta\ [\Sigma\{M^{jj'}\ \underline{x}^{j*} \otimes \underline{x}^{j'} \mid j,j' \in J_\alpha\}$$

$$= \Sigma\{M^{jj'}\ \pi_\beta^* \otimes \pi_\beta\ (\underline{x}^{j*} \otimes \underline{x}^{j'}) \mid j,j' \in J_\alpha\}$$

$$\pi_\beta^* \otimes \pi_\beta(U_\beta) = \Sigma\{M^{jj'}\ \pi_\beta^*\ (\underline{x}^{j*}) \otimes \pi_\beta\ (\underline{x}^{j'}) \mid j,j' \in J_\alpha\}$$

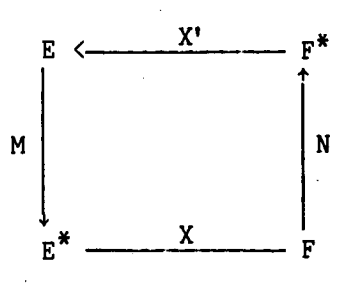
$$= \Sigma\{M^{jj'}\ \underline{x}_\beta^{j*} \otimes \underline{x}_\beta^{j'} \mid j,j' \in J_\alpha\}$$

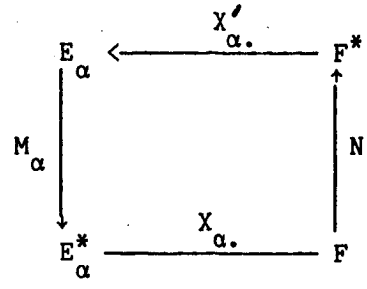
Comme $j,j' \in J_\alpha$, on a : $M^{jj'} = c_\alpha\ M_\alpha^{jj'}$ d'où :

$$\pi_\beta^* \otimes \pi_\beta\ (U_\beta) = c_\alpha\ \Sigma\{M_\alpha^{jj'}\ \underline{x}_\beta^{j*} \otimes \underline{x}_\beta^{j'} \mid j,j' \in J_\alpha\} = c_\alpha\ U_{\alpha\beta}$$

démonstration symétrique pour montrer que $\pi_\alpha^* \otimes \pi_\alpha (Z_\beta) = c'_\beta Z_{\alpha\beta}$.

On s'intéresse maintenant aux triplets $(X_{\alpha.}, M_\alpha, N)$ $\alpha \in A$ et (X, M, N) soient les tenseurs associés U_α, $U \in F^* \otimes F$. Les schémas de dualités sont :





En appliquant aux triplets précédents les relation (III.4.3.1) et (III.4.3.2), on a :

$$U = \Sigma \{c_\alpha U_\alpha \mid \alpha \in A\} \text{ et symétriquement } Z = \Sigma \{c'_\beta Z_\beta \mid \beta \in B\}$$

La linéarité de la trace permet d'avoir le corollaire suivant :

corollaire (III.4.3.1)

On a les relations suivantes entre les mesures d'informations associées aux différents tableaux.

$I(X_{.\beta}) = \Sigma \{c_\alpha I(X_{\alpha\beta}) \mid \alpha \in A\}$
$I(X_{\alpha.}) = \Sigma \{c'_\beta I(X_{\alpha\beta}) \mid \beta \in B\}$
$I(X) = \Sigma \{c_\alpha I(X_{\alpha.}) \mid \alpha \in A\} = \Sigma \{c'_\beta I(X_{.\beta}) \mid \beta \in B\}$
$I(X) = \Sigma \{c_\alpha c'_\beta I(X_{\alpha\beta}) \mid \alpha \in A, \beta \in B\}$

III.4.4 Réduction de l'information

Pour étudier l'ensemble des triplets $(X_{\alpha\beta}, M_\alpha, N_\beta)$, $\alpha \in A$, $\beta \in B$, on étudiera donc le triplet (X, M, N) en effet, on a d'après le corollaire III.4.3.1 :

$$I(X, M, N) = \Sigma \{c_\alpha c'_\beta I(X_{\alpha\beta}, M_\alpha, N_\beta) \mid \alpha \in A, \beta \in B\}$$

donc l'information I(X, M, N) est une "moyenne" pondérée des informations relatives à chaque tableau.

Le problème que l'on cherchera à résoudre est alors l'approximation de rang K du tenseur $X_{E \otimes F}$ relativement à la métrique $M \otimes N$ dans $E \otimes F$.

On pourra le voir sous plusieurs points de vue :

III.4.5 Etude du tableau : "histoire des variables"

On considère l'ensemble des triplets $(X_{\alpha.}, M_\alpha, N)$, $\alpha \in A$: le tableau X est partitionné en colonnes, c'est le cas par exemple si les tableaux $X_{\alpha.}$ sont des groupes de variables relatives à un même ensemble d'individus. Les sous-espaces vectoriels considérés sont (E_α, M_α), $\alpha \in A$, les espaces des individus relatifs aux groupes de variables, l'espace total est $E = \sum \{E_\alpha \mid \alpha \in A\}$ muni de la métrique $M = \Sigma \{c_\alpha M_\alpha \mid \alpha \in A\}$.

(F, N) : l'espace des variables. Les mesures d'informations associées aux tableaux vérifient $I(X) = \Sigma \{c_\alpha I(X_{\alpha.}) \mid \alpha \in A\}$ d'après le corollaire (III.4.3.1)

Le tableau X étudié est la juxtaposition des tableaux $X_{\alpha.}$ en ligne, c'est le tableau : "histoire des variables".

Si $T_{E \otimes F}$ est une approximation d'ordre K de $X_{E \otimes F}$, nous avons vu que si $F_K = T(E^*)$, espace image de E^* par T, et il existe une base $\{v_k \mid k \in K\}$ N-orthonormée tel que les vecteurs v_k soient le plus lié au sens de $\mathcal{L}$ au triplet (X, M, N).

Ainsi, le premier vecteur v_1 de l'ensemble est solution du problème :

Problème III.4.5.1

$$\max_{v_1 \in F} \mathcal{L}(c_1, X, M) = \langle U, v_1^* \otimes v_1 \rangle$$

$$\text{avec } \| v_1 \|^2_M = 1$$

comme $U = \Sigma \{c_\alpha U_\alpha \mid \alpha \in A\}$, on a :

$$< U, v_1^* \otimes v_1 > = \Sigma \{c_\alpha < U_\alpha, v_1^* \otimes v_1 > \mid \alpha \in A\}$$

$$= \Sigma \{c_\alpha \mathcal{L} (v_1, X_\alpha, M) \mid \alpha \in A\}$$

La démarche est donc analogue à celle de Caroll, on recherche une variable unique $v_1 \in F$, la plus liée aux différents paquets de variables X_α, $\alpha \in A$, non pas au sens de la liaison R^2 de corrélation multiple ni au sens de la liaison L^2 d'Escofier mais au sens de la liaison $\mathcal{L}$ que nous avons définie (produit scalaire du tenseur représentatif du triplet et le projecteur associée à v_1).

Cette liaison $\mathcal{L}$ est la plus générale, on verra que les liaisons R^2 et L^2 ne sont l'expression de cette liaison pour des métriques particulières.

Une fois, le vecteur v_1 obtenu, le processus étant réitéré, le vecteur v_2 est la variable unique orthogonale à v_1 la plus liée aux groupes X_α ets, ...

L'étude de la liaison entre plusieurs groupes de variables X_α. sera approfondie au paragraphe III.5.

III.4.6 Etude du tableau "histoire des individus"

On considère les triplets $(X_{.\beta}, M, N_\beta)$, $\beta \in B$. Le tableau X est partitionné en ligne, cela correspond par exemple à des tableaux $X_{.\beta}$ où des groupes d'individus ont été mesurés relativement au même ensemble de variables ont été fait à des instants différents. Les sous-espaces vectoriels considérés sont :

(E, M) l'espace des individus

(F_β, N_β), $\beta \in B$, les espaces de variables relatifs aux groupes d'individus, l'espace total est $F = \sum \{F_\beta, \mid \beta \in B\}$ muni de la métrique $N = \Sigma \{c_\beta N_\beta \mid \beta \in B\}$, les mesures d'informations associées aux tableaux vérifient : $I(X) = \Sigma \{c'_\beta I(_{.\beta}) \mid \beta \in B\}$ d'après le corollaire (III.4.3.1) le tableau X étudié ici est la juxtaposition des tableaux $X_{.\beta}$ en colonnes, c'est le tableau "histoire des individus".

Au tenseur $T_{E \otimes F}$, meilleure approximation d'ordre K de $X_{E \otimes F}$, correspond donc un ensemble de vecteurs $\{u_k \mid k \in K\}$ M-orthonormés les plus "liés" au sens de $\mathcal{L}$ des tableaux $X_{.\beta}$ (raisonnement identique qu'au paragraphe précédent). Du point de vue mathématique, en effet, les études des triplets (X_α, M_α, N), et $(X_{.\beta}, M, N_\beta)$ sont symétriques.

Ainsi, le premier vecteur $u_1 \in E$ est le vecteur général le plus lié aux différents tableaux $X_{.\beta}$, Z étant le tenseur de $E^* \otimes E$ relatif au triplet (X, M, N), le vecteur u_1 maximise :

$$< Z, u_1^* \otimes u_1 > = \Sigma \{c'_\beta \mathcal{L}(u_1, X_\beta, N_\beta) \mid \beta \in B\}$$

Pour interpréter la liaison $\mathcal{L}$, il faut remarquer qu'en Analyse des Données, les métriques N_β, relatives aux espaces F_β, seront toujours diagonales, ce sont des pondérations relatives aux individus (même constante, la plupart du temps). On posera donc $N_\beta = D_{p_\beta}$. On suppose que les matrices $X_{.\beta}$ sont centrées.

Si l'on note V_β, les matrices de variances-covariances relatives aux tableaux X_β, rappelons que : $Z_\beta = X'_\beta D_{p_\beta} X_\beta M = V_\beta M$ pour $\beta \in B$.

$$\mathcal{L}(u_1, X_\beta, D_{P_\beta}) = < Z_\beta, u_1^* \otimes u_1 > = \text{trace}\,(V_\beta M u_1 u_1' M)$$

$$= \text{trace}\,(u_1' M \dot{V}_\beta M_{u_1}) = < u_1, V_\beta M u_1 >_M$$

$$= I_{\overset{\perp}{\Delta u_1}} M^{(N_J^{I_\beta})} \quad (\text{car } \| u_1 \|_M^2 = 1)$$

où I_β est l'ensemble des individus relatif au tableau $X_{.\beta}$ ($I = U\{I_\beta \mid \beta \in B\}$) et $N_J^{I_\beta}$ est le nuage des individus dans E. La liaison $\mathcal{L}$ relative aux groupes d'individus s'interpréte donc comme l'inertie du nuage des individus relative à l'hyperplan $\overset{\perp M}{\Delta u_1}$ M-orthogonal à u_1. Ainsi, u_1 est le vecteur de E maximisant :

$$< Z, u_1^* \otimes u_1 > = \Sigma \{c'_\beta \mathcal{L}(u_1, X_\beta, D_{P_\beta}) \mid \beta \in B\}$$

$$= \Sigma \{c'_\beta I_{\overset{\perp}{\Delta u_1}} M^{(N_J^{I_\beta})} \mid \beta \in B\}$$

ou encore u_1 est le vecteur minimisant :

$$< Z, \bigotimes^{1} {}^{M} u_1 > = \Sigma \{c'_\beta \, I_{\Delta u_1} \, (N_J^{I_\beta}) \mid \beta \in B\}$$

ainsi, le vecteur u_1 est donc un vecteur d'inertie minimum en "moyenne" relatif aux groupes d'individus $X_{.\beta}$.

Une situation typique où les pondérations D_{p_β} diffèrent selon les groupes d'individus I_β (en fait des modalités) se rencontre par exemple, lorsque l'on traite un ensemble de tableaux de contingence juxtaposés en colonnes par l'analyse des correspondances.

III.4.7 Mesure d'information liée à un tableau croisé

Dans cette partie, nous nous intéressons, si l'on a un ensemble de triplets $(X_{\alpha\beta}, M_\alpha, N_\beta)$ $\alpha \in A$, $\beta \in B$ aux formes quadratiques

$$V^{\beta}_{\alpha\alpha} = X'_{\alpha\beta} N_\beta X_{\alpha\beta} \quad , \quad W^{\alpha}_{\beta\beta} = X_{\alpha\beta} M_\alpha X'_{\alpha\beta}$$

ou encore aux applications

$$V^{\beta}_{\alpha'\alpha} = X'_{\alpha'\beta} \, N_\beta \, X_{\alpha\beta} \in L(E^*_\alpha, E_{\alpha'}) \text{ et } W^{\alpha}_{\beta'\beta'} = X_{\alpha\beta} \, M_\alpha \, X'_{\alpha\beta} \in L(F^*_\beta, F_{\beta'})$$

et leurs généralisations à plusieurs triplets. Les tableaux qui leur sont associés jouent, en effet, en Analyse des Données un grand rôle.

Par exemple :

1) Si $(X, D_{1|P_x}, D_p)$ et $(Y, D_{1|P_y}, D_p)$ sont deux variables qualitatives $V = XD_pY$ est le tableau de probabilités associé au tableau de contingence, son analyse est équivalente à faire l'AFC de X et Y.

2) Si $(X_q, D_{1|P_q}, D_p)$ sont un ensemble de variables qualitatives en posant $X = \pi\{X_q \mid q \in Q\}$ l'on a $B = X D_p X'$ est le tableau de probabilité relatif

au tableau de Burt associé aux variables X_q. Si X_q est tableau de nombres positifs, alors B est le tableau de Burt généralisé proposé par Benzecri [Ben 82] et l'analyse du tableau de Burt généralisé ou non donne les mêmes facteurs que l'analyse du tableau X.

3) Si (X, M, D_p) est un tableau de variables quantitatives centrées, alors $V = XD_pX'$ est le tableau de corrélations ou de covariance suivant que les variables sont réduites ou non.

Nous étudierons les mesures d'information et opérateurs liés à ces tableaux. Les résultats fournis permettront de guider le choix des coefficients de pondérations $\{c_q \mid q \in Q\}$. Il est en effet connus, par exemple, qu'il est plus judicieux de choisir les coefficients $\{c_q \mid q \in Q\}$ relativement à l'étude d'un ensemble de variables qualitatives en raisonnant sur le tableau de Burt qu'à partir du tableau disjonctif complet (cf [Caz 80].

L'information (inertie) relative au tableau disjonctif complet, uniquement fonction du nombre de modalités de variables qualitatives est peu intéressante, en effet, comparée à celle du tableau de Burt associé.

Une autre application des résultats obtenus dans ce paragraphe est leurs utilisations dans le cadre de la classification automatique. La classification automatique peut, en effet, être présentée comme l'étude du tableau croisant la variable qualitative partition X_k et l'ensemble des tableaux $\{X_q, q \in Q\}$ (cf paragraphe IV.2.2).

Soit un triplet (X,M,N) et $V = X'NX \in L(E^*,E)$, $W = XMX' \in L(F^*,F)$ on a alors la proposition :

Proposition III.4.7.1

Les opérateurs associéss aux triplets (V,M,M) et (W,N,N) sont Z^2 et U^2 on a alors :

$$I(V) = I(W) = \| U \|^2 = \| Z \|^2$$

Démonstration :

Soit O_1, l'opérateur associé au triplet (V,M,M), son expression matricielle est :

$$O_1 = VMVM = Z^2$$

de même si O_2 désigne celui associé à (W,N,N) on a alors :

$$O_2 = WNWN = U^2$$

Par suite $I(V) = \text{trace } U^2 = \text{trace } (U.U) = \| U \|^2 = \Sigma\{\lambda_r^2 \mid r \in [1,\ldots,r_X]\}$ où r_X désigne le rang de X et λ_r la r-ème valeur propre de U, ces valeurs propres qui sont égaux à celles de Z d'où $I(V) = I(W)$.

Si on a $X = \pi\{X_q \mid q \in Q\}$ un ensemble de variables qualitatives, cette proposition exprime que l'analyse des correspondances de X et celui du tableau de Burt donne les même facteurs et si λ est valeur propre dans l'analyse de X, λ^2 est valeur propre de l'analyse du tableau de Burt.

Nous nous plaçons dans le cas où l'on considère un ensemble de triplets $(X_{\alpha\beta}, M_\alpha, N_\beta)$ $\alpha \in A$, $\beta \in B$ auxquels sont associés les opérateurs $U_{\alpha\beta}$ et $Z_{\alpha\beta}$, on note $V^\beta_{\alpha'\alpha} = X'_{\alpha'\beta} N_\beta X_{\alpha\beta}$ l'application linéaire de $L(E^*_\alpha, E_{\alpha'})$ auquel correspond le tenseur $V^\beta_{E_\alpha \otimes E_{\alpha'}}$ de $E_\alpha \otimes E_{\alpha'}$.

On considère alors les décompositions :

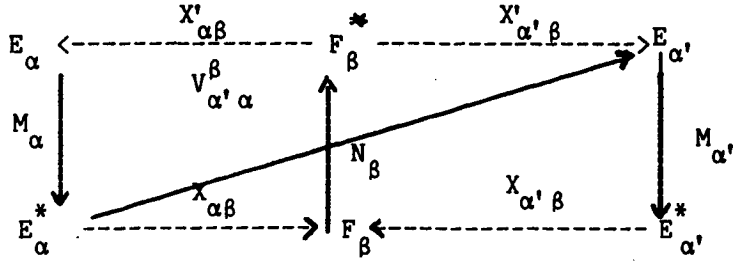
$$E = \Sigma\{E_\alpha \mid a \in A\} \text{ et } E \otimes E = \Sigma\{E_\alpha \otimes E_{\alpha'} \mid \alpha,\alpha' \in A\}$$

et l'on note

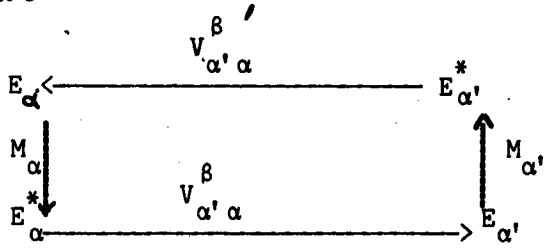
$V^\beta_{E \otimes E} = \pi\{V^\beta_{E_\alpha \otimes E_{\alpha'}} \mid \alpha,\alpha' \in A\}$, le tableau associé à $V^\beta_{E \otimes E}$ étant

$V^\beta = \pi\{V^\beta_{\alpha'\alpha} \mid \alpha,\alpha' \in A\}$ où $V^\beta_{\alpha'\alpha}$ est celui associé à $V^\beta_{E_\alpha \otimes E_{\alpha'}}$.

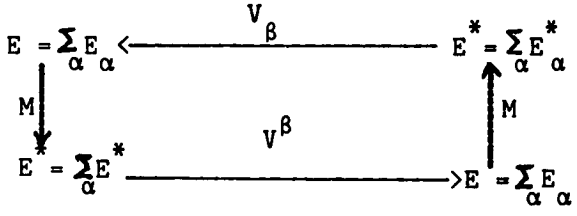
Les triplets considérés sont $(V^{\beta}_{\alpha'\alpha}, M_{\alpha}, M_{\alpha'})$ $\alpha,\alpha' \in A$, (V^{β}, M, M) où M est la métrique pondérée $M = \Sigma \{c_{\alpha} M_{\alpha} \mid \alpha \in A\}$ et les schémas de dualité associés sont :



ou encore



celui relatif à (V^{β}, M, M) étant



De manière symétrique on a :

$$W^{\alpha}_{\beta'\beta} = X_{\alpha\beta}\, M_{\alpha}\, X'_{\alpha\beta'}$$

le tenseur associé : $W^{\alpha}_{F_{\beta} \otimes F_{\beta'}}$ de $F_{\beta} \otimes F_{\beta'}$ les espaces somme directes considérés sont :

$$F = \Sigma \{F_{\beta_{\alpha}} \mid \beta \in B\}\ ,\ F \otimes F = \Sigma \{F_{\beta} \otimes F_{\beta'} \mid \beta,\ \beta' \in B\}$$

l'on pose :

$$W^{\alpha}_{F \otimes F} = \pi \{W^{\alpha}_{F_{\beta} \otimes F_{\beta'}} \mid \beta,\ \beta' \in B\}$$

le tableau associé à $W^{\alpha}_{F \otimes F}$ étant :

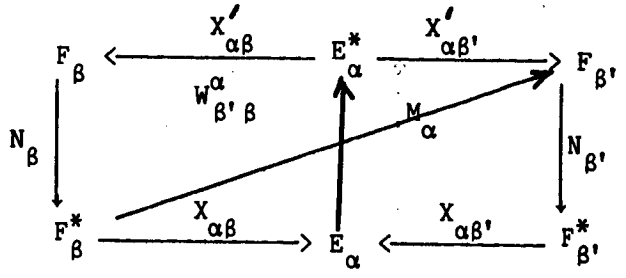
$$W^{\alpha} = \pi \{ W^{\alpha}_{\beta' \beta} \mid \beta, \beta' \in B \}$$

les triplets considérés sont :

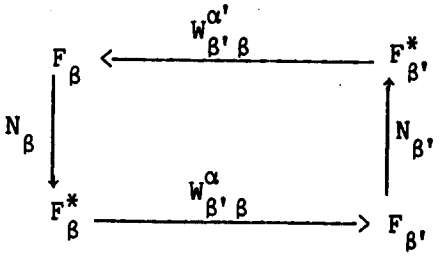
$$(W^{\alpha}_{\beta\beta}, N_{\beta}, N_{\beta'}) \text{ et } (W^{\alpha}, N, N)$$

où $N = \Sigma\{C_{\beta} N_{\beta} \mid \beta \in B\}$

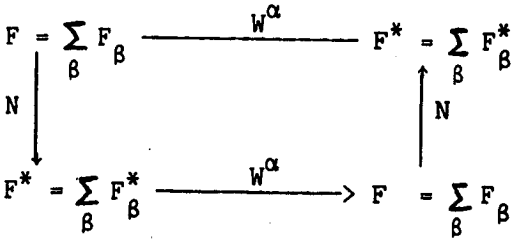
les schémas de dualité étant :



ou encore



et



soit f_{β} et e_{α} les tenseurs unités respectifs des espaces $F^{*}_{\beta} \otimes F_{\beta}$ et $E^{*}_{\alpha} \otimes E_{\alpha}$ on a alors :

Proposition III.4.7.2

On a les relations suivantes :

$$I(V^{\beta}_{\alpha'\alpha}) = \langle U_{\alpha\beta} , U_{\alpha'\beta} \rangle = \langle U_{\alpha\beta} \cdot U_{\alpha'\beta} , f_{\beta} \rangle$$

$$I(V^{\beta}) = \Sigma \{c_{\alpha} c_{\alpha'} \langle U_{\alpha\beta} , U_{\alpha'\beta} \rangle \mid \alpha, \alpha' \in A\}$$

$$= \| U_{\beta} \|^2$$

et symétriquement :

$$I(W^{\alpha}_{\beta\beta'}) = \langle Z_{\alpha\beta} , Z_{\alpha\beta'} \rangle = \langle Z_{\alpha\beta} \cdot Z_{\alpha\beta'} , e_{\alpha} \rangle$$

$$I(W^{\alpha}) = \Sigma\{c'_{\beta} c'_{\beta'} \langle Z_{\alpha\beta} , Z_{\alpha\beta'} \rangle \mid \beta, \beta' \in B\}$$

$$= \| Z_{\alpha} \|^2$$

Démonstration

Le tenseur $O^{\beta}_{\alpha'\alpha} \in E^*_{\alpha} \otimes E_{\alpha}$ associé au triplet $(V^{\beta}_{\alpha'\alpha}, M_{\alpha}, M_{\alpha'})$ a pour expression matricielle :

$$O^{\beta}_{\alpha'\alpha} = V^{\beta'}_{\alpha'\alpha} M_{\alpha'} \; V^{\beta}_{\alpha'\alpha} M_{\alpha}$$

on a donc :

$$I(V^{\beta}_{\alpha'\alpha}) = \text{trace } (O^{\beta}_{\alpha'\alpha}) = \text{trace } (V^{\beta'}_{\alpha'\alpha} M_{\alpha'} \; V^{\beta}_{\alpha'\alpha} M_{\alpha})$$

comme $V^{\beta}_{\alpha'\alpha} = X'_{\alpha'\beta} N_{\beta} \chi_{\alpha\beta}$ on a :

$$I(V^{\beta}_{\alpha'\alpha}) = \text{trace}\,(X'_{\alpha\beta}\, N_\beta\, X_{\alpha'\beta}\, M_{\alpha'}\, X_{\alpha'\beta}\, N_\beta\, X_{\alpha\beta}\, M_\alpha)$$

$$= \text{trace}\,(X'_{\alpha\beta}\, N_\beta\, U_{\alpha'\beta}\, X_{\alpha\beta}\, M_\alpha)$$

$$= \text{trace}\,(U_{\alpha'\beta}\, X_{\alpha\beta}\, M_\alpha\, X'_{\alpha\beta}\, N_\beta) = \text{trace}\,(U_{\alpha'\beta} \cdot U_{\alpha\beta})$$

$$= \langle U_{\alpha\beta}\, ,\, U_{\alpha'\beta} \rangle = \langle U_{\alpha\beta} \cdot U_{\alpha'\beta}\, ,\, f_\beta \rangle$$

Nous avons défini la liaison entre deux triplets $(X_{\alpha\beta}, M_\alpha, N_\beta)$ et $(X_{\alpha'\beta}, M_{\alpha'}, N_\beta)$ considéré ici comme deux paquets de variables portant sur même ensemble d'individus comme le produit scalaire entre les tenseurs associés $U_{\alpha\beta}$ et $U_{\alpha'\beta}$ de l'espace $F^*_\beta \otimes F_\beta$

$$\mathcal{I}\left[(X_{\alpha\beta}, M_\alpha)\, ,\, (X_{\alpha'\beta}, M_{\alpha'})\right] = \langle U_{\alpha\beta}\, ,\, U_{\alpha'\beta} \rangle = I(V^{\beta}_{\alpha'\alpha})$$

cette information est justement contenue dans celle du triplet $(V^{\beta}_{\alpha'\alpha}, M_\alpha, M_{\alpha'})$. Ce dernier triplet est donc bien représentatif pour l'étude des liaisons entre les triplets $(X_{\alpha\beta}, M_\alpha, N_\beta)$ et $(X_{\alpha'\beta}, M_{\alpha'}, N_\beta)$.

La deuxième égalité montre que l'on peut trouver des décompositions optimales de $I(V^{\beta}_{\alpha'\alpha})$ en cherchant un ensemble de vecteurs $\{v_k \mid k \in K\}$ de $F^*_\beta \otimes F_\beta$, en étudiant l'opérateur $U_{\alpha\beta} \cdot U_{\alpha'\beta} \in F^*_\beta \otimes F_\beta$, l'ensemble de vecteurs $\{v_k \mid k \in K\}$ N_β-orthonormés maximisant l'expression :

$$\langle u_{\alpha\beta} \cdot U_{\alpha'\beta}\, ,\, \Sigma\{v^*_k \otimes v_k \mid k \in K\} \rangle$$

ce qui revient à étudier les éléments propres de la matrice $U_{\alpha\beta} \cdot U_{\alpha'\beta} \cdot N_\beta$.

Lorsque $X_{\alpha\beta}$, $X_{\alpha'\beta}$ sont deux paquets de variables quantitatives et les métriques $M_\alpha = (V_{\alpha\alpha})^{-1}$, $M_{\alpha'} = (V_{\alpha'\alpha})^{-1}$ les métriques de Mahalanobis associés et $N_\beta = D_p = \frac{1}{n_\beta}\, Id_\beta$, on a alors $U_{\alpha\beta} = A_\alpha$, $U_{\alpha'\beta} = A_{\alpha'}$ les D_p-projecteurs $U_{\alpha\beta} \cdot U_{\alpha'\beta} = A_\alpha \cdot A_{\alpha'}$ on reconnait les équations de la canonique entre les paquets de variables $X_{\alpha\beta}$ et $X_{\alpha'\beta}$.

La relation $V^{\beta}_{E \otimes E} = \pi\{V^{\beta}_{E_\alpha \otimes E_{\alpha'}} \mid \alpha,\ \alpha' \in A\}$ entraîne d'après le lemme (III.4.2.1). :

$$\| V^{\beta}_{E \otimes E} \|^2 = \Sigma\{c_\alpha\, c_{\alpha'}\, \| V^{\beta}_{E_\alpha \otimes E_{\alpha'}} \|^2 M_\alpha \otimes M_{\alpha'} \mid \alpha,\ \alpha' \in A\}$$

$$= \Sigma\{ < c_\alpha\, c_{\alpha'} < U_{\alpha\beta},\ U_{\alpha'\beta} > \mid \alpha, \alpha'\ A\} = \| U_\beta \|^2$$

en effet $U_\beta = \Sigma\{c_\alpha\, U_{\alpha\beta} \mid \alpha \in A\}$ est le tenseur de $F^*_\beta \otimes F_\beta$ associé au triplet $(X_{.\beta}$, M, $N_\beta)$. On étudiera donc le triplet $(V^\beta$, M, M) pour étudier les liaisons entre les triplets $(X_{\alpha\beta}, M_\alpha, N_\beta)$ $\alpha \in A$.

De manière symétrique, on démontre les égalités relatives à $W^{\alpha}_{\beta'\,\beta}$ et W^α.

Soient $A_1 \subset A$, $A_2 \subset A$ deux sous-ensembles de A, on considère les décompositions :

$$E_{A_1} = \sum\{E_{\alpha_1} \mid \alpha_1 \in A_1\}\ ,\ E_{A_2} = \sum\{E_{\alpha_2} \mid \alpha_2 \in A_2\}$$

d'où

$$E_{A_1} \otimes E_{A_2} = \sum\{E_{\alpha_1} \otimes E_{\alpha_2} \mid \alpha_1 \in A_1,\ \alpha_2 \in A_2\}$$

et l'on considère

$$V^{\beta}_{E_{A_1} \otimes E_{A_2}} = \pi\ \{V^{\beta}_{E_{\alpha_1} \otimes E_{\alpha_2}} \mid \alpha_1 \in A_1,\ \alpha_2 \in A_2\},$$

le tableau correspondant étant $V^{\beta}_{A_1A_2}$, le triplet associé :

$(V^{\beta}_{A_1A_2}, M_{A_1}, M_{A_2})$ où $M_{A_1} = \Sigma\ \{c_{\alpha_1} M_{\alpha_1} \mid \alpha_1 \in A_1\}$ et $M_{A_2} = \Sigma\ \{c_{\alpha_2} M_{\alpha_2} \mid \alpha_2 \in A_2\}$

on a alors

Préposition III.4.7.3

Les égalités suivantes sont vérifiées :

$$I(V^{\beta}_{A_1A_2}) = \Sigma\{c_{\alpha_1} c_{\alpha_2} I(V^{\beta}_{\alpha_1\alpha_2}) \mid \alpha_1 \in A_1,\ \alpha_2 \in A_2\} = \Sigma\{c_{\alpha_1} c_{\alpha_2} \langle U_{\alpha_1\beta},\ U_{\alpha_2\beta}\rangle \mid \alpha_1 \in A_1,\ \alpha_2 \in A_2\}$$

Le résultat est une conséquence immédiate du lemmme relatif à la décomposition de produits tensoriels, on a en effet d'après le lemme III.4.2.1.

$$M_{A_1} \otimes M_{A_2} = \Sigma\{c_1 c_2 M_{\alpha_1} \otimes M_{\alpha_2} \mid \alpha_1 \in A_1,\ \alpha_2 \in A_2\}$$

par définition :

$$I(V^{\beta}_{A_1A_2}) = \| V^{\beta}_{E_{A_1} \otimes E_{A_2}} \|^2_{M_{A_1} \otimes M_{A_2}} = \Sigma\{c_{\alpha_1} c_{\alpha_2} \| V^{\beta}_{E_{\alpha_1} \otimes E_{\alpha_2}} \|^2_{M_{\alpha_1} \otimes M_{\alpha_2}} \mid \alpha_1 \in A_1,\ \alpha_2 \in A_2\}$$

$$= \Sigma\{c_{\alpha_1} c_{\alpha_2} I(V^{\beta}_{\alpha_1\alpha_2}) \mid \alpha_1 \in A_1,\ \alpha_2 \in A_2\} = \Sigma\{c_{\alpha_1} c_{\alpha_2} \langle U_{\alpha_1\beta},\ U_{\alpha_2\beta}\rangle \mid \alpha_1 \in A_1,\ \alpha_2 \in A_2\}$$

en posant :

$$U_{A_1\beta} = \Sigma\{c_{\alpha_1} U_{\alpha_1\beta} \mid \alpha_1 \in A_1\} \text{ et } U_{A_2\beta} = \Sigma\{c_{\alpha_2} U_{\alpha_2\beta} \mid \alpha_2 \in A_2\}$$

on a $I(V^{\beta}_{A_1A_2}) = \langle U_{A_1\beta},\ U_{A_2\beta}\rangle = \langle U_{A_1\beta} \cdot U_{A_2\beta},\ f_{\beta}\rangle$

En étudiant les éléments propres de $U_{A_1\beta} \cdot U_{A_2\beta} \cdot N_{\beta}$, on pourrait avoir des décompositions optimales de $I(V^{\beta}_{A_1A_2})$.

Exemples

1) Supposons que l'on étudie la liaison entre un triplet $(X_o, D_1, \sigma^2, D_p)$ et Q-triplets $(X_q, D_1, \sigma^2, D_p)$ où X_o, X_q, $q \in Q$ sont des ensembles de variables quantitatives centrées relatives à un même ensemble d'individus.

$$V_{oQ} = \pi\{c_q X'_o D_p X_q \mid q \in Q\} = \pi\{c_q V_{oQ} \mid q \in Q\}$$

où V_{oq} est la matrice de variance-covariance de X_o et x_q. On a :

$$U_o = \Sigma\left\{\frac{1}{\text{var}\underline{x}^{j_o}} \underline{x}^{j_o *} \otimes \underline{x}^{j_o} \mid j_o \in J_o\right\}$$

$$U_q = \Sigma \left\{ \frac{1}{\text{var}\underline{x}^{j_q}} \; \underline{x}^{j_q *} \otimes \underline{x}^{j_q} \mid j_q \in J_q \right\} , \; q \in Q$$

où J_o, J_q sont les ensembles d'indices relatifs à X_o et X_q, on a alors :

$$I(V_{oq}) = \Sigma \{c_q < U_o, U_q > \mid q \in Q\}$$

$$= \Sigma \left\{ c_q \frac{(< \underline{x}^{j_o}, \underline{x}^{j_q} >_{D_p})^2}{\text{var}\underline{x}^{j_o} \; \text{var}\underline{x}^{j_q}} \mid j_o \in J_o, \; j_q \in J_q, \; q \in Q \right\}$$

$$= \Sigma \left\{ c_q \frac{\text{covar}^2 (\underline{x}^{j_o}, \underline{x}^{j_q})}{\text{var}\underline{x}^{j_o} \; \text{var}\underline{x}^{j_q}} \mid j_o \in J_o, \; j_q \in J_q, \; q \in Q \right\}$$

$$= \Sigma \{c_q \; \text{corr}^2 (\underline{x}^{j_o}, \underline{x}^{j_q}) \mid j_o \in J_o, \; j_q \in J_q, \; q \in Q\}$$

La mesure d'information relative au tableau V_{oQ} est donc la somme des corrélations au carré entre la variable à expliquer et les variables explicatives.

2) Si les triplets $(X_o, D_1 \mid P_o, D_p)$ et $(X_q, D_1 \mid P_q, D_p)$ sont des ensembles de variables qualitatives munis de la métrique du chi-deux respectives, les opérateurs associés aux triplets sont les D_p- projecteurs A_o, $A_q \in F^* \otimes F$, de $X_o(E_o^*)$ et $X_q(E_q^*)$, $q \in Q$.

l'on a :

$$V_{oQ} = \pi \{c_q \; X_o' \; D_p \; X_q \mid q \in Q\} = \pi \{c_q \; P_{oq} \mid q \in Q\}$$

où P_{oq} est le tableau des probabilités d'association des modalités de X_o et X_q et l'on a :

$$I(V_{oq}) = \Sigma \{c_q < A_o, A_q > \mid q \in Q\} = \Sigma \{c_q \; \Phi^2_{oq} \mid q \in Q\} + \Sigma \{c_q \mid q \in Q\}$$

où Φ^2_{oq} est le phi-deux du tableau de probabilité P_{oq}.

3) De manière générale, en utilisant les diverses expressions que l'on a données au paragraphe (II.3.3) du produit scalaire entre deux tenseurs $< U_q, U_{q'} >$ nous retrouvons les liens classiques entre l'étude des liaisons des triplets (X_q, M_q, D_p) , $(X_{q'}, M_{q'}, D_p)$ et l'étude du triplet $(V_{qq'}, M_q, M_{q'})$ ainsi si X_q et $X_{q'}$ sont des variables qualitatives, l'étude de leur liaison revient à faire l'AFC du tableau de contingence associé, si X_q est une variable qualitative et $X_{q'}$ une variable quantitative, l'étude de leurs liaisons (Analyse discriminante) revient à faire l'ACP sur le tableau $V_{qq'}$ qui est la matrice des centres de gravité relatifs aux modalités de la variable qualitative, et l'étude de la liaison entre une variable qualitative X_o et un paquet de variables qualitatives X_q, $q \in Q$ revient à faire l'AFC sur le tableau $V_{oq} = \pi \{V_{oq} \mid q \in Q\}$ juxtaposition des tableaux de contingences etc, ...

III.5 ETUDE D'UN ENSEMBLE DE VARIABLES DEFINIES SUR LE MEME ENSEMBLE D'INDIVIDUS

III.5.1 Introdution

Nous abordons dans ce chapitre, les applications des résultats précédents dans le domaine de l'Analyse des Données.

Nous avons montré au paragraphe (III.4.6) que l'étude du tableau "histoire des individus" était simplifié en Analyse des Données car les métriques N_β relatives aux variables seront toujours diagonales et généralement constantes. L'étude du tableau "histoire des variables" est plus complexe car les métriques M_α relatives aux individus peuvent être quelconques.

Nous avons déjà montré que l'étude d'un tableau pouvait se ramené à la recherche d'un ensemble de variables générales $\{v_k \mid k \in K\}$ ordonnées de moins en moins liées au sens de $\mathcal{L}$ aux différents tableaux (cf paragraphe III.4.5) . Nous allons approfondir ce point de vue.

III.5.2 La liaison $\mathcal{L}$ entre plusieurs ensembles de variables

Nous reprenons ici les notations du chapitre II, nous supposons que nous avons :

$\{X_q \mid q \in Q\}$ un ensemble de variables portant sur un même ensemble d'individus I

$\{M_q \mid q \in Q\}$ un ensemble de métriques associés aux tableaux X_q

$N = D_p$ un ensemble de pondérations relatives aux individus.

Du point de vue espaces vectoriels, la situation se présente comme suit :

(E_q, M_q) $q \in Q$ les différents espaces et les métriques associées relatifs aux individus. On note $E = \sum \{E_q \mid q \in Q\}$ l'espace total des individus muni de la métrique pondérée $M = \Sigma \{c_q M_q \mid q \in Q\}$ où $\{c_q \mid q \in Q\}$ est un système de pondération relatif aux tableaux X_q. Nous supposerons dorénavant que le nuage des individus est centré.

(F,N) est l'espace des variables.

Dans un premier temps, nous nous limitons à la recherche du premier vecteur v le plus lié aux paquets de variables X_q. Nous définissons comme liaison d'ordre 1 entre Q paquets de variables $X_1, \ldots, X_Q$ la valeur du maximum du problème d'optimisation (III.5.2.1).

Définition III.5.2.1

La liaison entre Q paquets de variables X_q, $q \in Q$ est la solution du problème : III.5.2.1

$$\mathcal{L}(X_1, \ldots, X_q) = \max_{v \in F} \Sigma \{c_q \mathcal{L}(v, X_q, M_q) \mid q \in Q\}$$
$$\text{avec } \| v \|^2_{D_p} = 1$$

Nous avons vu que v solution du problème est tel que (cf III.3.3)

$$\mathcal{L}(X_1, \ldots, X_q) = \Sigma \{c_q < U_q, v^* \otimes v > \mid q \in Q\} = < U, v^* \otimes v >$$

où $U = \Sigma \{c_q U_q \mid q \in Q\}$ est le tenseur associé au triplet (X, M, N) avec $X = \pi \{X_q \mid q \in Q\}$.

Nous allons donner les expressions de $\mathcal{L}(X_1, \ldots, X_q)$ et la procédure pour résoudre le problème d'optimisation :

Nous avons vu au paragraphe (II.3.2) que U_q s'exprimait en fonction de ses éléments propres comme suit :

$$U_q = \Sigma \{\lambda_i^q \phi_i^{q*} \otimes \phi_i^q \mid i \in I\}$$

par suite $\mathcal{L}(v, X_q, M_q) = < \Sigma \{\lambda_i^q \phi_i^{q*} \otimes \phi_i^q \mid i \in I\}, v^* \otimes v >$

$$= \Sigma \{\lambda_i^q (<v, \phi_i^q>_{D_P})^2 \mid q \in Q\}$$

$$= \Sigma \{\lambda_i^q \cos^2 (v, \phi_i^q) \mid q \in Q\}$$

en effet v et les vecteurs ϕ_i^q sont de normes 1. Nous allons montrer au paragraphe (III.5.4) que v solution du problème est centrée. Par suite :

$$\boxed{\mathcal{L}(v, X_q, M_q) = \Sigma \{\lambda_i^q \text{ corr}^2 (v, \phi_i^q) \mid q \in Q\}}$$

on a alors :

$$\boxed{\mathcal{L}(X_1, \ldots, X_Q) = \Sigma \{c_q \lambda_i^q \text{ corr}^2 (v, \phi_i^q) \mid q \in Q, i \in I\}}$$

La démarche proposée revient donc à chercher un vecteur v le plus correlé avec un coefficient de pondération $c_q \lambda_i^q$ aux facteurs ϕ_i^q des différents paquets de variables X_q.

Nous allons donner d'autres expressions de $\mathcal{L}(X_1, \ldots, X_Q)$ permettant le calcul effectif de la variable v.

$$\mathcal{L}(v, X_q, M_q) = \text{trace}\,(W_q D_p\, vv' D_p) = \text{trace}\,(v' D_p\, W_q D_p v)$$

comme $v' D_p\, W_q D_p v$ est un scalaire

$$\mathcal{L}(v, X_q, M_q) = v' D_p\, W_q D_p v$$

d'où

$$\boxed{\mathcal{L}(v, X_q, M_q) = \langle v, W_q D_p v \rangle_{D_p}}$$

par suite :

$$\Sigma \{\mathcal{L}(v, X_q, M_q) \mid q \in Q\} = \Sigma \{c_q \langle v, W_q D_p V\rangle_{D_p} \mid q \in Q\}$$

$$= \langle v, [\Sigma \{c_q W_q D_p \mid q \in Q\}] (v) \rangle_{D_p}$$

$$= \langle v, WD_p\, v \rangle_{D_p}$$

on a donc à résoudre :

Problème III.5.2.2

$$\mathcal{L}(X_1, \ldots, X_q) = \max \langle v, WD_p\, v \rangle_{D_P}$$

$$v \in R^I$$

$$\text{avec } \| v \|_{D_p} = 1$$

d'où la proposition suivante :

Proposition III.5.2.1

La liaison $\mathcal{L}(X_1, \ldots, X_Q)$ est égale à la plus grande valeur propre de WD_p associée à la première composante principale c^1.

Montrons que les liaisons R^2 de Caroll et L^2 d'Escofier ne sont que l'expression de $\mathcal{I}$ pour des métriques particulières.

Proposition III.5.2.2

On a les relations suivantes :

$\mathcal{I}(v, X_q, V_q^{-1}) = R^2(v, X_q)$ corrélation multiple entre la variable v et les variables $\{\underline{x}^j \mid j \in J_q\}$ du groupe q

$\mathcal{I}(v, X_q, \Delta_q) = L^2(v, X_q) = I_{\Delta v}(N_I^q)$ inertie de la projection du nuage des variables N_I^q relatif au groupe q. Δ_q étant la métrique diagonale dont les termes diagonaux sont : $\{M^j \mid j \in J_q\}$

Démonstration

- Associons à chaque groupe de variables X_q, la métrique de Mahalanobis V_q^{-1}. Cela revient à donner à l'espace R^{J_q} une structure isotrope, tout vecteur est axe d'inertie. L'opérateur U_q associé au triplet (X_q, V_q^{-1}, D_p) à pour expression (cf II.3.3)

$$U_q = X_q V_q^{-1} X_q' D_p = A_q$$

c'est-à-dire l'opérateur de projection associé à l'espace E_q, par suite :

$$\mathcal{I}(v, X_q, V_q^{-1}) = \langle v, A_q v \rangle_{D_p} = R^2(v, X_q)$$

- Si l'on se limite aux métriques diagonales Δ_q

$$\mathcal{I}(v, X_q, \Delta_q) = \langle v, W_q D_p v \rangle_{D_p} = L^2(v, X_q)$$

qui s'interprète comme l'inertie en projection du nuage N_I^q des variables. Cela n'est pas vrai si M_q est une métrique quelconque. C'est une des raisons pour laquelle ces auteurs se limitent aux métriques diagonales.

La liaison que nous proposons généralise donc bien les liaisons R^2 de Caroll et L^2 d'Escofier-Pages.

Pour résoudre le problème d'optimisation (III.5.2.2), nous allons présenter l'Analyse en Composantes Généralisées qui permettra de trouver une solution en considérant l'espace $\mathbb{R}^J$ contenant les individus. Auparavant, nous rappelons quelques résultats classiques en ACP.

III.5.3 Rappels d'Analyse en Composantes Principales :

L'Analyse en Composantes Principales, les méthodes factorielles ont un double objectifs

1) Une description de l'ensemble des individus par la recherche d'une représentation euclidienne simple du nuage des individus. Cette réduction s'obtient en recherchant les axes d'inertie minimum et en projetant le nuage sur les plans formés par de tels axes.

2) La recherche de composantes principales, combinaisons linéaires des variables originales de variances maximum fournissant de "bons résumés" de l'information.

Nous allons exprimer les problèmes d'optimisation résolus en ACP à l'aide des tenseurs $U \in F^* \otimes F$ et $Z \in E^* \otimes E$ associés au triplet (X, M, D_p). On se place dans cette partie dans la situation classique où M est une métrique diagonale ($M = Id$ ou $M = D_{1/\sigma^2}$).

Proposition III.5.3.1

Le vecteur unitaire u_1 du premier axe principal du nuage N_J^I des individus est solution du problème

Problème III.5.3.1

$$\left|\begin{array}{l} \max < Z, u_1^* \otimes u_1 > = I_{\Delta u_1}(N_J^I) \\ u_1 \in \mathbb{R}^J \\ \text{avec } \| u_1 \|^2 = 1 \end{array}\right.$$

En effet $Z = \Sigma \{p_i \underline{x}_i^* \otimes x_i \mid i \in I\}$, $< Z, u_1^* \otimes u_1 > = \Sigma \{p_i (< \underline{x}_i, u_1 >_M)^2 \mid i \in I\}$ d'où $< Z, u_1^* \otimes u_1 > = I_{\Delta u_1}(N_J^I)$ qui est bien le critère optimisé en ACP. Du point de vue matricielle, nous avons vu (cf paragraphe III.4.6) que le problème (III.5.3.1) est équivalent à :

Problème III.5.3.2

$$\left|\begin{array}{l} \max < u_1, VMu_1 >_M \\ u_1 \in E \\ \text{avec } \| u_1 \|^2_M = 1 \end{array}\right.$$

De même symétriquement, on la situation suivante, relativement aux variables

Proposition III.5.3.2

Le vecteur unitaire v_1 relatif à la première composante principale est solution du problème

Problème III.5.3.3

$$\left|\begin{array}{l} \max <U, v_1^* \otimes v_1> = \Sigma \{M^j (<\underline{x}^j, v_1>_{D_p})^2 \mid j \in J\} = I_{\Delta v_1}(N_I^J) \\ v_1 \in F \\ \text{avec } \| v_1 \|^2 D_p = 1 \end{array}\right.$$

Matriciellement le problème se posent sous la forme suivante :

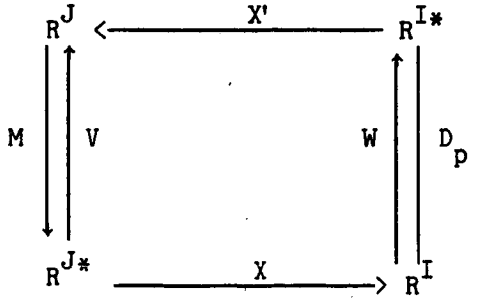
Problème III.5.3.4

$$\left|\begin{array}{l} \max \langle v_1, WD_p v_1 \rangle_{D_p} \\ v_1 \in F \\ \text{avec } \| v_1 \|^2_{D_p} = 1 \end{array}\right.$$

Les résultats classiques en ACP montrent que les matrices VM et WD_p ont même valeurs propres, il suffit donc d'étudier celle dont la dimension est la plus petite. En obtenant les K premiers vecteurs propres de VM, $\{v_k \mid k \in K\}$, cet ensemble est le système d'axes d'inertie minimum, des formules de passages simples ($v_k = XM\, u_k$, $k \in K$) permettent d'obtenir les composantes principales vecteurs propres de WD_p.

III.5.4 L'Analyse en Composantes Principales Généralisées :

La solution du problème d'optimisation relative à $(X_1, \ldots, X_Q)$ est donc donnée par l'ACP du triplet (X, M, D_p) où $X = \pi \{X_q \mid q \in Q\}$ et $M = \Sigma \{c_q M_q \mid q \in Q\}$, on diagonalisera VM puisque VM et WD_p ont même valeurs propres. Le schéma de dualité est :



on a alors la proposition suivante :

Proposition III.5.4.1

1) Le vecteur $v \in R^I$ de norme 1 maximisant $\Sigma\{\mathcal{J}(v,X_q,M_q) \mid q \in Q\}$ est homothétique à la première composante principale c^1 de l'ACP du triplet (X, M, D_p).

2)La valeur de ce maximun $\lambda = \mathcal{J}(X_1, \ldots, X_Q)$ est la plus grande valeur propre de la matrice VM où V désigne la matrice de variance-covariance $X' D_p X$.

L'expression de VM est si l'on note $V_{qq'}$ la matrice de variance-covariance des groupes X_q et $X_{q'}$

$$VM = \begin{bmatrix} V_{11} & V_{12} & \cdots & V_{1Q} \\ V_{21} & V_{22} & & \\ \vdots & & \ddots & \\ V_{Q1} & & & V_{QQ} \end{bmatrix} \begin{bmatrix} c_1 M_1 & & & \\ & & & 0 \\ & & \ddots & \\ & 0 & & c_Q M_Q \end{bmatrix}$$

- Lorsque l'on associe les métriques de Mahalanobis aux tableaux X_q, la matrice VM s'écrit :

$$VM = \begin{bmatrix} V_{11} & V_{12} & \cdots & V_{1Q} \\ V_{21} & V_{22} & & \\ \vdots & & \ddots & \\ V_{Q1} & & & V_{QQ} \end{bmatrix} \begin{bmatrix} V_{11}^{-1} & & & 0 \\ & V_{22}^{-1} & & \\ & & & \\ & 0 & & V_{QQ}^{-1} \end{bmatrix}$$

- La proposition permet de résoudre l'analyse canonique généralisée de Caroll en diagonalisant une matrice de dimension card J x card J.

- Ce sont SAPORTA-MASSON qui les premiers ont mis en évidence ces résultats dans le cas particulier ci-dessus. [cf [Sap 75] [Mas 74]].

- Si les variables sont qualitatives, en remplaçant chaque variable par le groupe de ses indicatrices non centrées alors V est le tableau de Burt et la métrique

$V_{qq}^{-1} = (X_q\ X'_q)^{-1}\ \forall\ q \in Q$ où $(X_q X'_q)^{-1}$ est la matrice diagonale des inverses des effectifs des modalités de la q-ème variable. Le problème résolu est celui des correspondances multiples.

- D'une manière générale selon les types de variables considérés et les métriques choisies, on retrouve les diverses méthodes factorielles (Analyse des correspondances simples, analyse discriminante, etc, ...).

- Nous avons justifié au paragraphe (III.4.4) l'analyse de tableaux de différents types quantitatifs ou qualitatifs avec des métriques pouvant différer suivant les groupes considérés à partir de la mesure d'information associée à un tableau. Dans cette partie, l'analyse peut être vue comme l'étude de :

$$\mathcal{I}(X_1, \ldots, X_Q) = \Sigma \{\gamma_i^q \text{ corr}^2 (v, \phi_i^q) \mid q \in Q,\ i \in I\}$$

c'est-à-dire à chercher un facteur v le plus corrélé aux facteurs ϕ_i^q des différents groupes, corrélations pondérées par un coefficient $\gamma_i^q = c_q\ \lambda_i^q$. Ce coefficient est le produit de deux termes :

λ_i^q qui tient compte de la variance de la composante principale i du groupe q. C'est une pondération des facteurs ϕ_i^q à l'intérieur des groupes. Si la métrique choisie pour un groupe est celle de Mahalanobis cela revient à accorder la même importance à chaque facteur ϕ_i^q ($\lambda_i^q = 1$ pour tout $i \in [1, \ldots, \text{rang } X_q]$

c_q : c'est une pondération inter-groupe qui reste à discuter pour homogénéiser l'influence des divers groupes.

Remarquons que si $r_q = \text{rang } X_q$ alors il n'existe que $r = \Sigma \{r_q \mid q \in Q\}$ termes non nuls dans l'expression $(X_1, \ldots, X_Q)$ associés à des valeurs propres $\lambda_i^q \neq 0$.

Cette analyse, nous allons voir, est identique à une ACP sur le tableau juxtaposé C des composantes principales associées à chaque groupe X_q.

Proposition III.5.4.1

L'analyse en composantes principales du triplet (X, M, D_p) est identique à celle de (C, Id_E, D_p) en particulier $\mathcal{I}(X_1, \ldots, X_Q) = \mathcal{I}(C_1, \ldots, C_Q)$ où C_q désigne le

tableau des composantes principales du tablelau X_q.

Démonstration :

Nous avons vu que, si l'on note Γ_q, le tenseur associé au triplet (C_q, Id_{E_q}, D_p) et U_q celui associé à (X_q, M_q, D_p) que $\Gamma_q = U_q$ en effet soient $\{c_I^q \mid i \in [1,\ldots,r_q]\}$ les composantes principales relatives à X_q, D_p-orthogonaux et de normes λ_i^q :

$$C_i^q = \lambda^{q\frac{1}{2}} \phi_i^q \text{ et } \Gamma_q = \Sigma \{\lambda_i^q \quad \phi_i^{q*} \otimes \phi_i^q \mid i \in I\} = U_q$$

en attribuant la valeur $\lambda_i^q = 0$ pour les vecteurs ϕ_i^q, $i > r_q$ ajoutés à l'ensemble $\{\phi_i^q \mid i \in [1, \ldots, r_q]\}$ pour former une base D_p-orthonormés de $\mathbb{R}^I$.

Par suite :

$$\mathcal{L}(X_1, \ldots, X_Q) = \Sigma \{\mathcal{L}(v, X_q, M_q) \mid q \in Q\} = \Sigma \{\mathcal{L}(v, C_q, Id_q) \mid q \in Q\}$$

$$\boxed{\mathcal{L}(X_1, \ldots, X_Q) = \mathcal{L}(C_1, \ldots, C_Q)}$$

Plus généralement, les triplets (X, M, D_p) et (C, Id, D_p) sont équivalents car :

$$U = \Sigma \{c_q U_q \mid q \in Q\} = \Sigma \{c_q \lambda_i^q \phi_i^{q*} \otimes \phi_i^q \mid i \in I, q \in Q\}$$

$$\Gamma = \Sigma \{c_q \Gamma_q \mid q \in Q\} = \Sigma \{c_q \lambda_i^q \phi_i^{q*} \otimes \phi_i^q \mid i \in I, q \in Q\}$$

par suite : $\| U - \Gamma \|^2 = 0$ CQFD

En remarquant que la métrique Id_{E_q} associée aux tableaux C_q est diagonale nous avons une interprétation de la liaison $\mathcal{L}$ en termes d'inertie.

$$\mathcal{L}(v, C_q, Id) = \Sigma \{(< C_i^q, v >_{D_p})^2 \mid i \in I\}$$

$$= I_{\frac{1}{\Delta v}} (N_I^{C_q}) \quad \text{inertie de la projection du nuage}$$

$N_I^{C_q}$ des composantes principales C_q sur le vecteur v. La liaison $\mathcal{L}(X_1, \ldots, X_Q)$ est donc la moyenne pondérée des inerties des nuages des composantes principales $N_I^{C_q}$ sur le facteur v le plus lié aux différents facteurs ϕ_i^q

$$\mathcal{L}(X_1, \ldots, X_Q) = \Sigma \{c_q \, I_{\frac{1}{\Delta v}} (N_I^{C_q}) \mid q \in Q\}$$

III.5.5 Expression de la liaison $\mathcal{L}$ en fonction des variables :

Nous avons montré que la liaison $\mathcal{L}(X_1, \ldots, X_Q)$ s'exprimait facilement en fonction des facteurs des tableaux X_q. L'étude des liaisons R^2 de Caroll et de L^2 d'Escofier ont montré que cette liaison pouvait s'exprimer selon les métriques considérées en fonction des variables $\underline{x}^j$ des tableaux X_q.

Dans cette partie, après avoir donné l'expression générale de $\mathcal{L}(X_1, \ldots, X_Q)$ nous envisageons quelques situations particulières : tableaux de mesures, et tableaux de modalités et donnons les expressions de $\mathcal{L}(X_1, \ldots, X_Q)$ dans chaque cas :

$$U_q = \Sigma \{M_q^{jj'} \, \underline{x}^{j*} \otimes \underline{x}^{j'} \mid j, j' \in J\}$$

par suite

$$\mathcal{L}(v, X_q, M_q) = <U_q, v^* \otimes v> = \Sigma \{M_q^{jj'} <\underline{x}^j, v>_{D_p} <\underline{x}^{j'}, v>_{D_p} \mid j, j' \in J_q\}$$

et pour v relatif à $\mathcal{L}(X_1, \ldots, X_Q)$, on a :

$$\mathcal{L}\ (X_1, \ldots, X_Q) = \Sigma\ \{c_q\ M_q^{jj'} < \underline{x}^j, v >_{D_p} < \underline{x}^{j'}, v >_{D_p} \mid j,j' \in J_q,\ q \in Q\}$$

Lorsque l'on a un sèul tableau, on reconnait le critère maximisé par l'ACP classique.

III.5.5.1 Etude d'un ensemble de tableaux de mesures :

Si les tableaux X_q sont des tableaux de mesures homogènes, on peut choisir comme métriques $M_q = Id \quad \forall\ q \in Q$ ce qui correspond à effectuer une ACP non normée sur le tableau $X = \pi\ \{X_q \mid q \in Q\}$, la liaison s'écrit alors :

$$\mathcal{L}\ (X_1, \ldots, X_Q) = \Sigma\ \{c_q\ \mathrm{cov}^2\ (\underline{x}^j, v) \mid j \in J_q,\ q \in Q\}$$

les variables v, $\{\underline{x}^j \mid j \in J\}$ sont en effet centrées donc :

$$< \underline{x}^j, v >_{D_p} = \mathrm{cov}\ (\underline{x}^j, v)$$

Si les tableaux sont hétérogènes, on le normalise en choisissant la métrique de Sebestien $M_q = D_{1/\sigma^2}$, on effectue alors une ACP normée sur le tableau X, les variables $\{\underline{x}^j \mid j \in J\}$ étant donc centrées réduites, on a alors :

$$\mathcal{L}\ (X_1, \ldots, X_Q) = \Sigma\ \{c_q\ \mathrm{corr}^2\ (\underline{x}^j, v) \mid j \in J_q,\ q \in Q\}$$

La première composante principale est donc une variable la "plus corrélée en moyenne" à l'ensemble des variables $\{\underline{x}^j \mid j \in J\}$ propriété soulignée et mise en évidence pour la première fois par MASSON [Mas 74].

III.5.5.2 Etude d'un ensemble de tableaux de modalités

Dans cette partie, nous considérons un ensemble de variables de natures qualitatives. Ce type de tableaux est étudié par l'Analyse des Correspondances Multiples.

On considère une variable qualitative comme un tableau X_q constitué par l'ensemble de ses variables indicatrices.

On note traditionnellement :

$$f_{IJ} = \{f_{ij} = \frac{x_i^j}{k} \mid i \in I,\ j \in J\} \text{ où } x_i^j \in [0,1]$$

où k=n * card Q

le tableau de fréquence des couples (i,j)

$f_J = \{f_j = \Sigma \{f_{ij} \mid i \in I\}$ l'ensemble des poids associés aux modalités

$f_I^j = \{f_i^j = \frac{f_{ij}}{f_j} \mid i \in I\}$ le profil d'une modalité j dans l'espace $F \simeq \mathbb{R}^I$

$f_I^{J_q} = \{f_I^j \mid j \in J_q\}$ le nuage des modalités relatif à la variable q

$f_I = \{f_i = \frac{1}{n} \mid i \in I\}$ l'ensemble des pondérations relatives aux individus comme f_i est constant, on considérera l'espace $F \simeq \mathbb{R}^I$ comme muni de la métrique identité Id_I.

On a alors l'égalité des liaisons que nous avons définie, R^2 de corrélation multiple et L^2 d'Escofier. En effet la liaison L^2 a été construite comme une généralisation de la liaison R^2 de Caroll identique pour les variables qualitatives. Vérifions le rapidement.

Soit un vecteur ϕ^I de $\mathbb{R}^I$ et W_q l'espace engendré par les modalités f_I^j pour $j \in J_q$ et A_q le projecteur associé à W_q.

$$R^2(\phi^I, W_q) = \langle \phi^I, A_q \phi^I \rangle = \| A_q \phi^I \|^2$$

or $A_q = \Sigma \left\{ \frac{1}{\| f_I^j \|^2} f_I^{j*} \otimes f_I^j \mid j \in J_q \right\}$ car les profils f_I^j sont orthogonaux

$$\| f_I^j \|^2 = \Sigma \left\{ \frac{f_{ij}^2}{f_j^2} \mid i \in I \right\} = \frac{1}{k.f_j} \quad \text{les } f_{ij} \text{ étant égaux à 0 ou } \frac{1}{k}$$

par suite $A_q \, \phi^I = \Sigma \{ f_{jk} < f_I^j, \phi^I > f_I^j \mid j \in J_q \}$

$$\| A_q \, \phi^I \|^2 = \Sigma \{ f_j^2 k^2 (< f_I^j, \phi^I >)^2 \| f_I^j \|^2 \mid j \in J_q \}$$

$$= \Sigma \{ f_j k (< f_I^j, \phi^I >)^2 \mid j \in J_q \} = I_{\Delta_{\phi^I}}(N_I^{J_q}).k$$

inertie de la projection du nuage des modalités de J_q sur Δ_{ϕ^I}, donc

$$\boxed{\mathcal{I} (\phi^I, f_I^{J_q}, Id) = R^2 (\phi^I, W_q) = L^2(\phi^I, f_I^{J_q}).k}$$

Le vecteur ϕ^I de norme 1 maximisant

$$\Sigma \{ c_q \mathcal{I} (\phi^I, f_I^{J_q}, Id) \mid q \in Q \} = \Sigma \{ c_q < \phi^I, A_q \, \phi^I > \mid q \in Q \}$$

$$= < \phi^I, \Sigma \{ c_q A_q (\phi^I) \mid q \in Q \} >$$

est le vecteur propre associé à la plus grande valeur propre de $\Sigma\{c_q A_q \mid q \in Q\}$ La liaison entre les Q questions $J_1, \ldots, J_Q$ s'écrit donc en fonction de ϕ^I

$$\mathcal{I} (J_1, \ldots, J_Q) = \Sigma \{ c_q R^2 (\phi^I, W_q) \mid q \in Q \} = \Sigma \{ c_q \cos^2 (\phi^I, A_q \phi^I) \mid q \in Q \}$$

$$\boxed{\mathcal{I} (J_1, \ldots, J_Q) = \Sigma \{ c_q \cos^2 (\phi^I, f_I^j) \mid j \in J_q, q \in Q \}}$$

Soit ψ^J le facteur relatif à λ dans R^J, la formule de transition s'écrit :

$$\phi^I = \frac{1}{\sqrt{\lambda}} \Sigma \{\psi_j f_I^j \mid j \in J\}$$

Soit le vecteur canonique ξ_q projection de ϕ^I sur W_q, on a :

$$\xi_q = A_q(\phi^I) = \Sigma \left\{ < \phi^I, f_I^j > \frac{f_I^j}{\| f_I^j \|^2} \mid j \in J_q \right\}$$

en remplaçant ϕ^I par son expression et en remarquant que les termes non nuls dans l'expression $<\phi^I, f_I^j>$ sont sont ceux relatifs aux composants $j \in J_q$, on a :

$$\xi_q = \frac{1}{\sqrt{\lambda}} \{\psi_j <f_I^j, f_I^j> \frac{f_I^j}{\| f_I^j \|^2} \mid j \in J_q\}$$

d'où

$$\boxed{\xi_q = \frac{1}{\sqrt{\lambda}} \{\psi_j f_I^j \mid j \in J_q\}}$$

Ainsi donc à une constante multiplicative près $\frac{1}{\sqrt{\lambda}}$ la composante du vecteur canonique ξ_q sur les modalités f_I^j, $j \in J_q$ est la restriction du facteur ψ^I à l'ensemble J_q. Ce qui peut encore s'écrire :

$$\frac{<\phi^I, f_I^j>}{\| f_I^j \|^2} = \frac{\psi_j}{\sqrt{\lambda}} \quad \text{pour } j \in J.$$

III.5.5.3 Etude d'un tableau mixte

Nous allons exprimer la liaison $\mathcal{I}$ dans le cas où l'on a un ensemble de triplets que l'on considérera comme composé de trois types de triplets.

Un premier ensemble $(X_{q_1}, V^{-1}_{q_1 q_1}, D_p)$, $q_1 \in Q_1$ de tableaux quantitatifs, $(X_{q_2}, D_{1|\sigma^2_{q_2}}, D_p)$ un autre groupe de variables quantitatives $q_2 \in Q_2$; $(X_3, D_{1|P_{q_3}}, D_p)$ un ensemble de variables qualitatives $q_3 \in Q_3$.

La liaison entre ces tripets s'écrit alors si $Q = Q_1 \cup Q_2 \cup Q_3$:

$$\mathcal{L}(X_1, \ldots, X_Q) = \Sigma\{c_{q_1} R^2(v, X_{q_1}) \mid q_1 \in Q_1\} + \Sigma\{c_{q_2} \text{corr}^2(\underline{x}^j, v) \mid j \in J_{q_2}, q_2 \in Q_2\}$$
$$+ \Sigma\{c_{q_3} \cos^2(f_I^j, v) \mid j \in J_{q_3}, q_3 \in Q_3\}$$

III.5.6 Etude des pondérations des tableaux

Nous discutons dans ce paragraphe du problème du choix des coefficients $\{c_q \mid q \in Q\}$. Plusieurs solutions ont été proposées pour le choix des coefficients $\{c_q \mid q \in Q\}$ suivant les critères optimisés (cf [Caz 80]), [Sap 79]).

Ces études concernent en général des tableaux de même nature : quantitatif ou qualitatif, pour l'étude de tableaux mixtes, Escofier propose une pondération que l'on rappelera. Dans cette partie, nous présentons les principales solutions dans le cadre de notre méthodologie en nous basant sur les expressions de l'information I(X) et la liaison $\mathcal{L}$.

Rappelons que :

$$I(X) = \Sigma \{c_q I(X_q) \mid q \in Q\}, \text{ on notera dans la suite } I_q = I(X_q)$$

Remarquons que si l'on multiplie les c_q par une constante, I(X) se trouve multiplié par cette constante, on imposera donc une condition de normalisation : $\Sigma \{cq \mid q \in Q\} = 1\}$

La liaison $\mathcal{L}$ a pour expression :

$$\mathcal{L}(X_1, \ldots, X_Q) = \Sigma \{c_q \mathcal{L}(v, X_q, M_q) \mid q \in Q\}$$
$$= \Sigma \{c_q \lambda_i^q \text{corr}^2 (\phi_i^q, v) \mid i \in [1, \ldots, r_q] \mid q \in Q\}$$

si on note r_q le rang de X_q.

Ainsi $\mathcal{L}(X_1, \ldots, X_Q)$ est une somme de $r = \Sigma \{r_q \mid q \in Q\}$ de corrélations au carré c'est-à-dire de nombres compris entre 0 et 1 pondérés par les valeurs propres et les coefficients c_q. Un tableau $X_{q'}$ de rang $r_{q'} > r_q$, comportera donc plus de termes que le tableau X_q, de rang r_q, mais sera plus difficile à résumer par un vecteur unique v. D'autre part les valeurs propres peuvent être supérieures à 1 pour des tableaux de mesures traités par l'ACP, mais inférieur à 1 pour des tableaux de modalités traités par l'ACM. Ces diverses considérations montrent la nécessité d'un choix de pondérations. Les principales possibilités sont :

1) Choix de pondérations égales

Cette solution est adoptée si par exemple l'ensemble est constitué de tableaux que l'on pense être "homogènes" ou que l'on a aucune idée à priori.

On choisit alors $\{c_q = \frac{1}{Q} \mid q \in Q\}$. Ainsi, souvent, un ensemble de tableaux de modalités est traité ainsi par l'ACM classique.

2) Rendre égale la mesure d'information associée à chaque groupe

C'est un choix logique dans le cadre de notre démarche, les méthodes d'Analyse de Données étant des techniques de réduction de l'information I(X) moyenne des informations I_q des groupes. Les coefficients seront choisis alors comme solutions du système d'équations

Système III.5.6.1

$$\begin{cases} c_1 I_1 = c_2 I_2 = \ldots = c_Q I_Q \\ \Sigma \{c_q \mid q \in Q\} = 1 \end{cases}$$

on trouve facilement :

$$c_q = \frac{1}{\Sigma \{ \frac{I_q}{I_{q'}} \mid q' \in Q\}} = \frac{1/I_q}{\Sigma\{1/I_{q'} \mid q' \in Q\}} \quad , \; q \in Q$$

La métrique relative aux variables D_p étant diagonale, I_q s'interpréte comme l'inertie du nuage des individus $N_{J_q}^{I}$ relatif aux variables de X_q. Nous avons vu au paragraphe (III.2.2) que l'information relative à une variable centrée $\underline{x}^j$: $I(\underline{x}^j, Id, D_p) = var(\underline{x}^j)$, cette normalisation est donc analogue à celle où lorsque l'on étudie un tableau de mesures hétérogènes, on réduit les variables pour que les distances entre individus soient indépendantes des unités de mesures.

3) Rendre égal le plus grand moment d'inertie de chaque groupe

Lorsque M est une métrique diagonale, I_q s'interpréte comme l'inertie du nuage des variables $N_I^{J_q}$ et $\mathcal{I}(v, X_q, M_q) = I_{\Delta_v^{-1}}(N_I^{J_q})$ En notant que les nuages "allongés" ont plus d'influence que les nuages "ronds", dans la détermination des composantes (inertie de la projection du nuage plus importante), Escofier propose de normaliser l'influence des différents tableaux en égalisant le plus grand moment d'inertie de chaque tableau. Les coefficients sont alors : $c_q = \frac{1}{\lambda_1^q}$, $q \in Q$.

4) Rendre égal l'information "utile" de chaque groupe

On peut ne pas tenir compte de toute l'information I_q relatif à un tableau mais une approximation d'ordre $s_q < r_q$. Le nombre s_q étant par exemple déterminé par une étude préalable du tableau X_q. Ainsi les coefficients c_q sont déterminés par le système d'équations :

Système III.5.6.2

$$\begin{cases} c_1 I_1^{s_1} = c_2 I_2^{s_2} = \ldots = c_Q I_Q^{s_Q} \\ \Sigma \{c_q \mid q \in Q\} = 1 \end{cases}$$

où $I_q^{s_q} = \Sigma \{\lambda_i^q \mid i \in [1, \ldots, s_q]\}$

Si $s_q = 1$, $q \in Q$ on retrouve le choix d'Escofier et si $s_q = r_q$, la normalisation par I_q. Cette solution intermédiaire permet de tenir compte de l'importance des groupes non par une procédure automatique mais par un choix de l'utilisateur après analyse séparée préalable de chaque groupe.

Si on note $C_q^{s_q}$ la solution adoptée relative à un tableau q, notons k la valeur commune

$c_{q'}\ I_{q'}^{s_{q'}} = k$, pour $q' \neq q$, $q' \in Q$ on a alors :

$$C_q^{s_q} = \frac{k}{I_q^{s_q}} \text{ donc pour } s_q < s'_q \text{ c-à-d } I_q^{s_q} < I_q^{s'_q} \text{ alors } C_q^{s'_q} < C_q^{s_q}$$

plus on choisit d'axes pour déterminer $I_q^{s_q}$ plus le coefficient $C_q^{s_q}$ diminue donc l'importance accordée au tableau X_q diminue.

On a donc $C_q^{r_q} \leq C_q^{s_q} \leq C_q^1$ pour $1 \leq s_q \leq r_q$. Choisir la solution d'Escofier pour un groupe X_q, c'est lui accorder l'importance maximale qu'elle puisse avoir dans notre contexte et normaliser par la mesure d'information I_q c'est lui accorder l'importance minimale.

5) Choisir les coefficients de l'opérateur "compromis"

Dans son approche Escoufier propose de visualiser les objets en utilisant les éléments propres de l'opérateur "compromis" c'est-à-dire, sous la contrainte : $\{\Sigma\ c_q^2 = 1 \mid q \in Q\}$, l'opérateur $U = \Sigma\ \{c_q\ U_q \mid q \in Q\}$ combinaison linéaire de U_q de norme maximum. Cet opérateur "résume" au mieux les divers opérateurs associés aux tableaux et Saporta [Sap 79] considère qu'elle est optimale en ce sens qu'elle aboutit à une AFC, lorsque les variables sont de types qualitatives, dont la somme des carrés de toutes les valeurs propres est maximale. La représentation des objets se fait alors dans un cadre tenant compte des liaisons globales entre variables. La solution est, en effet, la première composante principale du tableau de produits scalaires des opérateurs $T = \{T_{qq'} = \langle U_q, U_{q'} \rangle \mid q, q' \in Q\}$ Lorsque les variables sont qualitatives, T est le tableau de ϕ^2 de Pearson ou des coefficients T^2 de Tschuprow (cf travaux de Saporta [Sap 79]).

Dans le cadre de notre approche, cela revient à étudier le triplet (V, M, M) où $V = X'MX$ est la forme quadratique définie sur E^* associée à X. Nous avons vu en effet au paragraphe (III.4.7) que le tenseur associé au triplet (V, M, M) est le tenseur U^2 et que la mesure de l'information $I(V) = \| U \|^2$ tient compte des diverses liaisons entre les triplets (X_q, M_q, D_p), $q \in Q$. Il est donc naturel de

chercher les coefficients $\{c_q \mid q \in Q\}$ qui maximisent l'information I(V) c'est-à-dire les solutions du problème :

Problème III.5.6.1

$$\left|\begin{array}{l} \max I(V) = \| U \|^2 = \Sigma \{\lambda_\ell^2 \mid \ell \in [1, \ldots, r_u]\} = \\ \qquad = \Sigma \{c_q c_{q'} \langle U_q, U_{q'} \rangle \mid q, q' \quad Q\} \\ \{c_q \mid c_q \in R^+ , q \in Q\} \\ \text{avec } \Sigma \{c_q^2 \mid q \in Q\} = 1 \end{array}\right.$$

où $\{\lambda_\ell \mid \ell \in [1, \ldots, r_u]\}$ sont les valeurs propres de U de rang r_U.

De tels coefficients $\{c_q \mid q \in Q\}$ tiendront compte des liaisons mutuelles entre les triplets, ce sont les composantes de la première composante principale de l'ACP du tableau des produits scalaires des opérateurs.

6) Pondération dans l'étude d'une variable qualitative à expliquer et un ensemble de variables explicatives

Pour étudier, dans une optique de discrimination, la liaison entre une variable qualitative à expliquer et un ensemble de variables qualitatives explicatives, il est courant de faire une AFC sur la juxtaposition des tableaux de contingence croisant la variable à expliquer et les variables explicatives. Cette analyse n'est pas optimale et n'est une vraie analyse discriminante que si les variables sont deux à deux indépendantes. Saporta propose alors [Sap 79] d'améliorer l'AFC en choisissant des coefficients qui optimisent le pouvoir discriminant des facteurs en maximisant la somme des valeurs propres.

On trouve alors que les coefficients sont les phi-deux à un coefficient près, entre la variable à expliquer et les variables explicatives.

En présentant le problème dans notre contexte, nous retrouvons ce résultat en le généralisant facilement à un ensemble de variables explicatives, quantitatives ou qualitatives.

Nous supposons que l'on a un triplet $(X_o, D_{1|P_o}, D_p)$ relatif à la variable qualitative à expliquer et un ensemble de variables explicatives que l'on partagera en deux groupes :

- le premier groupe :

$(X_{q_1}, D_{1|P_{q_1}}, D_p)$ un ensemble de variables qualitatives $q_1 \in Q_1$

- le deuxième groupe :

$(X_{q_2}, (V_{q_2q_2})^{-1}, D_p)$ un ensemble de variables quantitatives $q_2 \in Q_2$,

on pose $Q = Q_1 \cup Q_2$, étant donnée la nature des métriques choisies, chi-deux pour les variables qualitatives : $D_{1|P_{q_1}}$, et mahalanobis $(V_{q_2q_2})^{-1}$ pour celles quantitatives, les tenseurs associés sont les D_p-projecteurs $A_q \in F^* \otimes F$ sur $X_q(E_q^*)$, $q \in Q$.

Le tableau $V_{oQ} = \pi \{X'_o D_p X_q \mid q \in Q\}$ contient l'information mutuelle entre X_o et $\{X_q \mid q \in Q\}$. On étudie le triplet (V_{oQ}, M, D_p) où $M = \Sigma \{c_q M_q \mid q \in Q\}$ avec $M_q = D_{1|P_q}$ si $q \in Q_1$ et $M_q = (V_{qq})^{-1}$ si $q \in Q_2$.

Les coefficients $\{c_q \mid q \in Q\}$ seront choisis solutions du problème

Problème III.5.6.2

$$\max I(V_{oQ}) = \langle A_o, \Sigma \{c_q A_q \mid q \in Q\rangle = \Sigma \{c_q \langle A_o, A_q\rangle \mid q \in Q\}$$

$$c_q \in R^+, \; q \in Q$$

$$\text{avec} \quad \Sigma\{c_q^2 \mid q \in Q\} = 1$$

Soit $K = \inf (\dim E_o, \dim E_q)$ $q \in Q$, en général K sera le nombre de modalités de la variable à expliquer X_o. Désignons par $\{u_k, k \in K\}$ les vecteurs propres de $\Sigma \{c_q A_o A_q \mid q \in Q\}$ de normes 1

On a : $I(V_{oQ}) = \Sigma \{c_q \text{ trace } A_o A_q \mid q \in Q\}$

$= \Sigma \{c_q u'_k A_o A_q u_k \mid k \in K, q \in Q\}$

le problème (III.5.6.2) est équivalent à

Problème III.5.6.3

$$\left|\begin{array}{l} \max I(V_{oQ}) = \Sigma \{c_q u'_k A_o A_q u_k \mid k \in K, q \in Q\} \\ c_q \in R^+, \ q \in Q \\ \langle u_k, u_{k'} \rangle = \delta^k_{k'} ; \ k,k' \in K \\ \Sigma\{c^2_q \mid q \in Q\} = 1 \end{array}\right.$$

En utilisant les multiplicateurs de Lagrange $\{\lambda_k \mid k \in K\}$ et μ on trouve en dérivant par rapport à c_q l'expression :

$$W = \Sigma \{c_q u'_k A_o A_k u_k \mid k \in K, q \in Q\} - \Sigma \{\lambda_k u'_k u_k \mid k \in K\} - \mu(\Sigma \{c^2_q \mid q \in Q\}-1)$$

que c_q doit être proportionnel à $\Sigma \{u'_k A_o A_q u_k \mid k \in K\} = \langle A_o, A_q \rangle$, ainsi si X_q est une variable qualitative on retrouve bien :

$$c_q = \langle A_o, A_q \rangle = \phi^2_{oq} \text{ à un coefficient près}$$

et si X_q est une variable quantitative :

$$c_q = \langle A_o, A_q \rangle = \text{trace } (V^{-1}_{qq} B_q) = I(N^{G_q}_{E_q})$$

où B_q la matrice d'inertie "inter-classe" relative à X_q et c_q est donc l'inertie du nuage des centres de gravité $N^{G_q}_{E_q}$ dans l'espace E_q muni de la métrique Mahalanobis V^{-1}_{qq}.

Chaque tableau X_q est donc pondéré suivant l'importance de la liaison du tableau X_q avec le tableau X_o au sens de la liaison .

7) Pondération par la norme des opérateurs

En utilisant l'expression de $\mathcal{L}(X_1, \ldots, X_Q)$ en fonction des opérateurs U_q :

$$\mathcal{L}(X_1, \ldots, X_Q) = \Sigma\{c_q < U_q, v^* \otimes v> \mid q \in Q\}$$

comme $||v^* \otimes v||^2 = 1$, on peut choisir $c_q = \frac{1}{||U_q||}$, $\mathcal{L}(X_1, \ldots, X_Q)$ s'écrit alors :

$$\mathcal{L}(X_1, \ldots, X_Q) = \Sigma\{R_v(U_q, v) \mid q \in Q\}$$

comme une somme de termes compris entre 0 et 1. En l'absence d'idée précise, on choisira cette pondération ou celle où les c_q sont égaux à 1.

III.5.7 Généralisation

Dans la définition de $\mathcal{L}(X_1, \ldots, X_Q)$ donnée au paragraphe III.5.2.1, nous nous dommes limités pour des raisons de clarté à la recherche d'un seul vecteur v_1 unitaire maximisant l'expression :

$$\Sigma\{c_q \mathcal{L}(v_1, X_q, M_q) \mid q \in Q\}$$

il est évident que ce vecteur v_1 obtenu, on peut réitérer le processus en cherchant un deuxième vecteur v_2 maximisant la même expression sous contrainte d'orthogonalité au premier et ainsi de suite. D'après les propriétés d'optimalité de l'Analyse en Composantes Principales Généralisées du triplet (X, M, D_p), on obtient pour une liaison d'ordre K, les K-premiers facteurs $\{v_k \mid k \in K\}$ d'une telle analyse, on pose alors :

$$\mathcal{L}(X_1, \ldots, X_Q) = \Sigma\{c_q \mathcal{L}(v_k, X_q, M_q) \mid q \in Q, k \in K\}$$

et on en déduit facilement les différentes expressions de $\mathcal{L}(X_1, \ldots, X_Q)$.

III.5.8 Conclusions

Au terme de cette étude, nous avons donné une formulation générale unique en termes de réduction de l'information d'une famille de méthodes d'Analyse de Données traitant un ensemble de tableaux de données en pondérant chaque groupe de variables par des coefficients positifs. Nous avons étudié cette réduction de l'information sous leurs aspects géométriques et en termes de liaison entre variables. Nous avons explicité dans chaque cas, les problèmes d'optimisation associés. Il apparaît ainsi qu'aux différents choix usuels en Analyse des Données : choix de l'ensemble d'individus, choix de l'ensemble des variables s'ajoute lorsque l'on traite un ensemble de variables, le choix de coefficients équilibrant le rôle joué par les différents groupes de variables. Les étapes que nous proposons pour étudier de tels tableaux sont alors :

1) Etude préalable de chaque groupe de variables. Nous avons vu en effet que les éléments principaux (valeurs propres, vecteurs propres) de chaque groupe influent dans la détermination du ou des facteurs les plus liés aux différents groupes.

2) Normalisation des groupes de variables : deux types de normalisations sont à considérer : une normalisation intra qui consiste au choix de métrique sur l'ensemble des individus, cette métrique s'interprétant en termes de pondération des variables du groupe. Une normalisation inter qui consiste aux choix des coefficients de pondérations inter-groupes.

Ces coefficients seront choisis en fonction des objectifs poursuivis tel que cela a été présenté au paragraphe précédent.

3) Réduction de l'information : c'est la résolution du problème d'optimisation par les techniques d'Analyse des Données. Nous avons envisagé jusqu'à présent que les techniques de type factoriel au chapitre suivant nous considérons les méthodes de classification.

4) Interprétation des résultats : les interprétations, en restant uniquement dans le cadre du problème d'optimisation, se font dans l'optique traditionnelle des méthodes factorielles à l'occurence l'Analyse en Composantes Principales, c'est-à-dire on considère les corrélations et contributions des variables ou individus par groupe aux facteurs. Dans une optique de visualisation

des différents nuages, les données étant normalisées (la métrique M est l'identité), il est alors possible d'appliquer les techniques, proposées par divers auteurs, comme l'Analyse Factorielle Multiple d'Escofier ou STATIS.

- Une telle approche où les problèmes de normalisation, d'optimisation et d'interprétation sont séparés suit les recommandations de MODULAD [Mod 82] dans l'écriture de programmes informatiques relatifs à une méthode d'A.D.. Elle correspond par ailleurs à la philosophie du système SICLA [Ral 84] dont l'architecture reflète un tel découpage.

- Une autre possibilité pour traiter un tableau n-aire est de considérer le triplet des composantes principales (C_q, Id_E, D_p) en effet, on a vu (cf proposition III.5.4.1.) que ce triplet est équivalent à (X_q, M_q, D_p) pour $q \in Q$. Ce qui correspond à la pratique classique de remplacer un ensemble de variables quantitatives ou qualitatives par leurs composantes principales pour l'analyser par la classification automatique par exemple. Toutefois, si les composantes principales n'ont pas une signification claire, de telles analyses posent des problèmes d'interprétation.

IV ETUDE D'UN TABLEAU N-AIRE PAR LA CLASSIFICATION AUTOMATIQUE

IV.1 Introduction

Nous allons, dans ce chapitre, considérer les applications des résultats généraux concernant les tableaux n-aires du chapitre III à la classification automatique. Nous généralisons un ensemble de méthodes de classification à l'étude de tels tableaux. Les méthodes factorielles classiques supposent la population homogène, pour chercher des axes d'inerties minimum ou des composantes principales. Les méthodes de classification type MND ont pour objectif la recherche de classes homogènes selon des critères mesurant l'adéquation de la classe à sa "représentation". Suivant le type de représentation, on aboutit à des méthodes différentes. En particulier, lorsque la représentation d'une classe est une variété affine, les travaux de DIDAY [DID 75], concernant l'Analyse Factorielle Typologique, ont montré que, les K-variétés affines locales de la partition optimale ont, en général, un taux d'inertie supérieur à celle de l'analyse globale.

Dans l'étude de tableaux n-aires, nous examinons les différents cas où la représentation d'une classe est le centre de gravité de la classe et, lorsqu'elle est une variété affine. Nous faisons ensuite le lien entre ces méthodes, les Nuées Dynamiques Généralisées, l'Analyse Factorielle Typologique Généralisée et l'Analyse Canonique Généralisée Typologique qui est la recherche de liaisons locales entre variables.

IV.2 Nuées dynamiques généralisées

Nous supposons que nous avons un ensemble de triplets (X_q, M_q, D_p), $q \in Q$ de variables relatives à un même ensemble d'individus. La classification que nous envisageons ici est d'abord celle pour laquelle la représentation d'une classe est le centre de gravité de la classe. Nous allons la considérer sous cinq points de vue différents.

IV.2.1 Classification d'un ensemble d'individus par un ensemble de variables

Le cadre de référence est l'espace $E = \sum \{E_Q \mid q \in Q\}$ muni de la métrique pondérée $M = \Sigma \{c_q M_q \mid q \in Q\}$. Un individu $\underline{x}_i$ de E est repéré par Q composantes $\underline{x}_{iq}$:

$$\underline{x}_i = \pi \{\underline{x}_{iq} \mid q \in Q\} \text{ où } \underline{x}_{iq} = \pi_q (\underline{x}_i) \in E_q.$$

A une partition P à K classes sont associés, K centres de gravité $\underline{g}_k$ des classes P_k, $k \in K$ tels que :

$$\underline{g}_k = \pi \{\underline{g}_{kq} \mid q \in Q\} = \Sigma \{\frac{p_i}{p_k} \underline{x}_i \mid i \in I_k\}$$

on a immédiatement : $\underline{g}_{kq} = \pi_q(\underline{g}_k) = \Sigma \{\frac{p_i}{p_k} \pi_q (\underline{x}_i) \mid i \in I_k\}$

$$= \Sigma \{\frac{p_i}{p_k} \underline{x}_{iq} \mid i \in I_k\}$$

On notera $D_{pk} = \{\frac{p_i}{p_k} \mid i \in I_k\}$ l'ensemble des poids relatifs à la classe P_k.

Donc si $N_E^{I_k}$ est le nuage des individus relatifs à la classe P_k, $\underline{g}_{kq}$ est le centre de gravité du nuage $N_{E_q}^{I_k}$ projection du nuage $N_E^{I_k}$ dans E_q.

Calculons la distance d'un individu $\underline{x}_i$ au centre de gravité $\underline{g}_k$ de la classe P_k :

$$d^2_M (\underline{x}_i, \underline{g}_k) = \Sigma \{c_q \, d^2_{M_q} (\underline{x}_{iq}, \underline{g}_{kq}) \mid q \in Q\}$$

d'après la définition de la métrique M pondérée. Si on note alors $X = \pi\{X_q \mid q \in Q\}$ et $X^k = \pi \{X^k_q \mid q \in Q\}$ le tableau général et celui relatif à la classe P_k, on définit les applications inerties relatives à un point de E comme suit :

$$I_r : P(X) \times E \times M \longrightarrow \mathbb{R}^+$$

$$(A, g, M) \longrightarrow I(A, g, M) = \Sigma \{p_i \, d^2_M (\underline{x}_i, g) \mid \underline{x}_i \in A\}$$

On a aussi Q applications inerties I^q_r relatives aux espaces E_q

$$I^q_r : P(X^q) \times E_Q \times M_q \longrightarrow \mathbb{R}^+$$

$$(A^q, \underline{g}_q, M_q) \longrightarrow I_r(A^q, \underline{g}_q, M_q) = \Sigma\{p_i \, d^2_{M_q} (\underline{x}_{iq}, \underline{g}_q) \mid \underline{x}_{iq} \in A^q\}$$

par suite

$$I_r(X^k, \underline{g}_k, M) = \Sigma \{p_i \, d^2_M (\underline{x}_i, \underline{g}_k) \mid \underline{x}_i \in X^k\}$$

$$= \Sigma \{c_q \, p_i \, d^2_{M_q} (\underline{x}_{iq}, \underline{g}_{kq}) \mid \underline{x}_{iq} \in X^k_q, \; q \in Q\}$$

$$= \Sigma \{c_q \, I^q_r (X^q_k, \underline{g}_{kq}, M_q) \mid q \in Q\}$$

L'inertie du nuage $N^{I_k}_E$ relatif aux individus de la classe P_k est donc la moyenne pondérée de celle de $N^{I_k E}_{E_q}$. Il est donc licite de chercher dans l'espace E une partition P à K classes minimisant l'inertie intra :

$$\text{Iintra} = \Sigma \{I_r (X^k, \underline{g}_k, M) \mid k \in K\}$$

Le problème qui demeure est celui du choix des coefficients $\{c_q \mid q \in Q\}$ pondèrant les inerties des nuages $N_{E_q}^{I_k}$. Nous avons déjà discuté de ce problème dans le cadre de la liaison entre Q ensembles de variables. Nous la reformulons au paragraphe (IV.2.6) dans le contexte de la classification automatique. Nous allons présenter le deuxième point de vue.

IV.2.2 Recherche d'une variable qualitative liée à un ensemble de variables au sens de $\mathcal{L}$

D'après la relation classique : Inertie totale = Inertie intra + Inertie inter la méthode de classification considérée maximise l'inertie inter dont nous allons donner quelques expressions. On appelle G le tableau des centres de gravité à K lignes et J colonnes :

$$G' = \pi \{\underline{g}_k \mid k \in K\}$$

On note $D_{P_K} = \{p_k \mid k \in K\}$ l'ensemble des poids des k classes associées aux vecteurs $\underline{g}_k$ tel que $p_k = \Sigma \{p_i \mid i \in I_k\}$ pour $k \in K$.

Le triplet relatif aux centres de gravité est (G, M, D_{P_K}) et le tenseur associé est noté $B \in E^* \otimes E$ qui a pour expression :

$$B = \Sigma \{p_k \, g_k^* \otimes g_k \mid k \in K\}$$

La mesure d'information associée au triplet (G, M, D_{P_K}) est notée $I(G, M, D_{P_K})$ et correspond à l'inertie inter, en effet, on a :

$$\boxed{I(G, M, D_{P_K}) = I_r(N_E^G) = \Sigma \{p_k \| \underline{g}_k \|^2_M \mid k \in K\} = \langle B, e \rangle = \text{Iinter}}$$

Une partition peut être considérée comme une variable qualitative à K modalités donc caractérisée par un triplet $(X_k, D_{1|P_K}, D_p)$. Nous allons montrer que la méthode de classification revient à chercher, un triplet $(X_k, D_{1|P_K}, D_p)$ la plus liée au sens de la liaison $\mathcal{L}$ aux Q triplets (X_q, M_q, D_p).

Nous avons vu au paragraphe (III.4.7) que cette liaison est donnée par la mesure d'information relative au triplet $(V_{KQ}, M, D_{1|P_K})$. Montrons que ce triplet est équivalent à celui des centres de gravité (G, M, D_{P_k}).

En effet, on a $V_{KQ} = \pi \{X'_K, D_p, X_q \mid q \in Q\}$. Le $k^{ème}$ vecteur ligne de V_{KQ} $\underline{y}_k$, est la moyenne des individus $\underline{x}_i$ du tableau $X = \pi \{X_q \mid q \in Q\}$ pour la classe k : $\underline{y}_k = \Sigma \{p_i \underline{x}_i \mid i \in I_k\} = p_k \underline{g}_k$ et ceci pour $k \in K$.

Le tenseur O associé à $(V_{KQ}, M, D_{1|P_K})$ de l'espace $E^* \otimes E$ a pour expression : $O = \Sigma \{\frac{1}{p_k} p_k \underline{g}^*_k \otimes p_k \underline{g}_k \mid k \in K\} = \Sigma \{p_k g^*_k \otimes g_k \mid k \in K\}$, on a donc bien $O = B$. La liaison entre X_K et $\{X_q \mid q \in Q\}$ est donnée alors par la mesure d'information relative au triplet des centres de gravité (G, M, D_{P_K}) c'est-à-dire par l'inertie inter et l'on a :

$$I(G, M, D_{P_K}) = \mathcal{L}\,[(X_K, D_{1|P_K}) \mid \{(X_q, M_q), q \in Q\}] = \Sigma \{c_q < A_K, U_q > \mid q \in Q\}$$

où A_K, U_q désignent les tenseurs de $F^* \otimes F$ relatifs aux triplets $(X_K, D_{1|P_K}, D_p)$ et (X_q, M_q, D_p). On a alors la proposition :

Proposition IV.2.2.1

Une méthode de classification maximisant l'inertie inter, recherche donc une variable qualitative X_k la plus liée à l'ensemble des variables X_q au sens de $\mathcal{L}$. La liaison ayant pour expression :

$$\mathcal{L}\,[(X_k, D_{1|p_k}) \mid \{(X_q, M_q), q \in Q\}] = \Sigma \{c_q <A_k, U_q> \mid q \in Q\}$$

Cette expression nous l'avons déjà rencontrée et commentée pour quelques expressions particulières de U_q et M_q (cf paragraphe II.3.3).

Ainsi par exemple si on a des triplet $(X_K, D_{1|P_K}, D_p)$ relatifs à un ensemble de variables qualitatives X_q, on a vu que :

$$Iinter = I(G, M, D_{P_K}) = \Sigma\{c_q \langle A_K, A_q\rangle \mid q \in Q\} = \Sigma\{c_q \Phi^2_{Kq} \mid q \in Q\} + \Sigma\{c_q \mid q \in Q\}$$

où Φ^2_{Kq} est le phi-deux relatif au tableau de probabilités P_{Kq} des associations des modalités des variables X_K et X_q.

Dans le cas général $\langle A_q, U_q \rangle = I(N^G_{E_q})$ est l'inertie du nuage des centres de gravité dans l'espace E_q.

IV.2.3 Décomposition optimale de l'information relative aux triplets

Nous allons donner une interprétation de la méthode de classification comme la recherche d'une approximation d'ordre K de l'information relative aux tableaux X_q, $q \in Q$. Ce qui intuitivement se comprend très bien, en effet, en adoptant comme mode de représentation d'une classe son centre de gravité, maximiser l'inertie inter, c'est chercher une réduction optimale de l'information, relative à l'ensemble des tableaux, X_q concentrée dans les centres de gravité.

Soit $X = \pi \{X_q \mid q \in Q\}$ le tableau total et le tenseur $U \in F^* \otimes F$ associé au triplet (X, M, D_p) $q \in Q$ est tel que :

$$U = \Sigma \{c_q U_q \mid q \in Q\}$$

on a donc :

$$Iinter = I(G, M, D_{P_K}) = \Sigma \{c_q \langle A_K, U_q \rangle \mid q \in Q\} = \langle A_K, U \rangle$$

A_K est l'opérateur associé au triplet $(X_K, D_{1|P_K}, D_p)$ a pour expression

$$A_K = \Sigma \left\{ \frac{1}{P_k} \underline{x}^{*k} \otimes \underline{x}^k \mid k \in K \right\}$$

or $\underline{x}^k$ est la variable indicatrice numéro k de X_K, on a donc :

$$\underline{x}^k = \Sigma \{x^k_i \underline{f}_i \mid i \in I\} = \Sigma \{\underline{f}_i \mid i \in I_k\}$$

en posant $v_k = \frac{\underline{x}^k}{\sqrt{p_k}}$ on remarque que l'ensemble $\{v_k \mid k \in K\}$ est D_p-orthonormé en effet : $\| v_k \|^2_{D_p} = 1$ et si $k \neq k'$

$$D_p(\underline{x}^k, \underline{x}^{k'})) = \Sigma \{D_p (f_i, f_{i'}) \mid i \in I_k, i' \in I_{k'}\} = 0 \text{ car } I_k \cap I_{k'} = 0$$

on a donc bien $< v_k, v_{k'} > = \delta^k_{k'}$ par suite :

$$\text{Iinter} = I(G, M, D_{P_K}) = < U, \Sigma \{v^*_k \otimes v_k \mid k \in K\} >$$

expression que nous avions déjà étudié (cf III.3.3).

Ainsi donc maximiser l'inertie inter revient donc à chercher un ensemble de vecteurs $\{v_k \mid k \in K\}$ assurant une décomposition optimale de la mesure d'information $I(X, M, D_p)$ en imposant aux vecteurs $\{v_k \mid k \in K\}$ éléments de l'espace F d'être le K variables indicatrices d'une variable qualitative X_K.

Nous avons vu (cf III.3.3) que l'analyse en composantes principales du triplet (X, M, D_p) donne aussi une décomposition optimale d'ordre K de l'information $I(X, M, D_p)$, aucune contrainte n'étant imposée aux vecteurs v_k D_p-orthonormés. On peut donc considérer que la méthode de classification est une ACP particulière. Cette propriété a été mise en évidence dans un autre contexte par Lermann [Ler 79] et Govaert [Gov 83]. Ces auteurs raisonnent en termes "d'axes" et "inerties" et sont donc obligés de supposer que la métrique M soit diagonale (ou rendue diagonale après transformation des données). La notion "d'inertie" ne nous semble pas appropriée et même peut prêter à confusion lorsque l'on s'intéresse aux variables. En effet le nuage des variables n'est pas centré et les composantes principales ne sont pas des axes d'inertie minimum. Notre approche nous semble plus naturelle puisque nous n'avons besoin de faire aucune hypothèse sur la métrique M et raisonnons en termes de liaison et, d'autre part nous nuançons le résultat. En effet, l'ACP donne un ensemble de vecteurs $\{v_k \mid k \in K\}$ "ordonnés" dont chacune est solution d'un problème d'optimisation et dont le pouvoir de liaison vu en décroissant, ce qui n'est pas le cas pour les variables indicatrices d'une variable qualitative. C'est la différence entre les problèmes (III.3.3.1) et (III.3.3.4).

Pour illustrer la similitude entre l'ACP et la classification automatique considérons un ensemble de triplets de tableaux quantitatifs centrés :

$(X_K, D_{1|\sigma_q^2}, D_p)$, $q \in Q$. Soient $\{\underline{c}^k \mid k \in K\}$ l'ensemble des composantes principales de l'analyse du tableau $X = \pi \{X_q \mid q \in Q\}$ (ACP normé).

Les opérateurs U_q relatifs aux triplets, ont pour expression, en remarquant que $\text{var}\, \underline{x}^j = \| \underline{x}^j \|^2_{D_p}$:

$$U_q = \Sigma \left\{ \frac{1}{\| \underline{x}^j \|^2_{D_p}} \underline{x}^{j*} \otimes \underline{x}^j \mid j \in J_q \right\}$$

Les composantes principales optimisent le critère suivant :

$$W_1 = \Sigma \{c_q < U_q, \Sigma\{ \frac{\underline{c}^{k*} \otimes \underline{c}^k}{\| \underline{c}^k \|^2_{D_p}} \mid k \in K\} > \mid q \in Q\}$$

$$W_1 = \Sigma \{c_q < \Sigma\{\frac{\underline{x}^{j*} \otimes \underline{x}^j}{\| \underline{x}^j \|^2_{D_p}} \mid j \in J_q\}, \Sigma\{\frac{\underline{c}^{k*} \otimes \underline{c}^k}{\| \underline{c}^k \|^2_{D_p}} \mid k \in K\} > \mid q \in Q\}$$

$$W_1 = \Sigma \{c_q \frac{(<\underline{x}^j, \underline{c}^k>_{D_p})^2}{\| \underline{x}^j \|^2_{D_p} \| \underline{c}^k \|^2_{D_p}} \mid k \in K, j \in J_q, q \in Q\}$$

Celui optimisé par la classification automatique est :

$$W_2 = \Sigma \{c_q < U_q, \{\frac{\underline{x}^{k*} \otimes \underline{x}^k}{\| \underline{x}^k \|^2_{D_p}} \mid k \in K \} > \mid q \in Q\}$$

$$W_2 = \Sigma \{c_q \frac{(<\underline{x}^j, \underline{x}^k>_{D_p})^2}{\| \underline{x}^j \|^2_{D_p} \| \underline{x}^k \|^2_{D_p}} \mid k \in K, j \in J_q, q \in Q\}$$

On voit que les critère W_1 et W_2 sont identiques. La différence réside essentiellement sur le fait qu'en ACP les vecters $\{c^k \mid k \in K\}$ sont des variables numériques centrées, W_1 s'interpréte comme une somme de corrélations tandis qu'en classification automatique $\underline{x}^k$ est une variable indicatrice relative à la variable qualitative partition X_k, W_2 s'interprète comme une somme de cosinus.

IV.2.4 Approximation d'ordre k du tenseur relatif aux triplets

Nous avons montré au paragraphe III.3.2 que cette décomposition optimale recherchée dans $F^* \otimes F$ s'interprétait comme une approximation d'ordre K du tenseur $X_{E \otimes F}$ dans l'espace $E \otimes F$ muni de la métrique produit $M \otimes D_p$. Ainsi donc la classification automatique revient à déterminer la meilleure approximation : le tenseur $X_{E \otimes F_K}$ projection de $X_{E \otimes F}$ sur $E \otimes F_K$ tel que :

$\| X_{E \otimes F_K} \|^2_{M \otimes D_p} = \Sigma \{ < U, v_k^* \otimes v_k > \mid k \in K \}$ soit maximale F_k étant l'espace vectoriel engendré par les variables indicatrices $\{ v_k \mid k \in K \}$ de la variable X_K.

IV.2.5 Recherche d'une métrique optimisant l'information totale

Nous allons montrer que la méthode de classification proposée revient à chercher une certaine métrique relative à l'espace R^I : N(P) dépendant de la partition P telle que la mesure d'information $I(X, M, N(P))$ soit maximale.

Soit la métrique N_k définit sur $\mathbb{R}^{I_k}$ telle que :

$$N_k = \{ N_k^{ii'} = \frac{p_i \, p_{i'}}{p_k \, p_k} \mid i, i' \in I_k \}$$

on a alors le lemme :

Lemme IV.2.5.1

Les égalités suivantes sont vérifiées :

$$p_k \| g_k \|^2_M = < U, v_k^* \otimes v_k > = p_k \, I(X^k, M, N_k)$$

Le schéma de dualité étant :

$$\begin{array}{ccc} E \simeq \mathbb{R}^J & \xleftarrow{\quad X^{k'} \quad} & \mathbb{R}^{I_k *} \\ M \downarrow & & \downarrow N_k \\ \mathbb{R}^{J*} & \xrightarrow{\quad X^k \quad} & \mathbb{R}^{I_k} \end{array}$$

Démonstration :

calculons $p_k \| g_k \|^2_M = p_k < \Sigma \{ \frac{p_i}{p_k} \underline{x}_i \mid i \in I_k \}, \Sigma \{ \frac{p_{i'}}{p_k} \underline{x}_{i'} \mid i' \in I_k \} >_M$

$$= \Sigma \{ \frac{p_i \, p_{i'}}{p_k} < \underline{x}_i, \underline{x}_{i'} >_M \mid i, i' \in I_k \}$$

$$= \Sigma \{ M^{jj'} \frac{p_i \, p_{i'}}{p_k} x_i^j \, x_{i'}^{j'} \mid i, i' \in I_k \; ; \; j, j' \in J \}$$

On note $f_i^{*D_p} = < . , f_i >_{D_p}$ le D_p-projecteur sur f_i pour le distinguer du vecteur f_i^* de la base duale. On a alors :

$$v_k^* \otimes v_k = \left(\frac{1}{\sqrt{p_k}} \Sigma \{ f_{i'}^{*D_p} \mid i' \in I_k \} \right) \otimes \left(\frac{1}{\sqrt{p_k}} \Sigma \{ f_i \mid i \in I_k \} \right)$$

$$= \frac{1}{p_k} \Sigma \{ f_{i'}^{*D_p} \otimes f_i \mid i, i' \in I_k \}$$

par suite :

$$<U, v_k^* \otimes v_k> = <\Sigma \{ M^{jj'} \underline{x}^{*j} \otimes \underline{x}^{j'} \mid j, j' \in J \}, \Sigma \{ \frac{1}{p_k} f_{i'}^{*D_p} \otimes f_i \mid i, i' \in I_k \}>$$

$$= \Sigma \{ M^{jj'} \frac{1}{p_k} < \underline{x}^{*j} \otimes \underline{x}^{j'}, f_{i'}^{*D_p} \otimes f_i > \mid j, j' \in J \; ; \; i, i' \in I_k \}$$

$$= \Sigma \{ M^{jj'} \frac{1}{p_k} < \underline{x}^j, f_i >_{D_p} < \underline{x}^{j'}, f_{i'} >_{D_p} \mid j, j' \in J \; ; \; i, i' \in I_k \}$$

$$= \Sigma \{ M^{jj'} \frac{p_i \, p_{i'}}{p_k} x_i^j \, x_{i'}^{j'} \mid j, j' \in J \; ; \; i, i' \in I_k \}$$

d'où $p_k \| \underline{g}_k \|^2_M = < U, v_k^* \otimes v_k >$ l'examen de ces expressions montrent que si l'on pose $N_k^{ii'} = \frac{p_i \, p_{i'}}{p_k p_k}$ l'expression s'écrit :

$$< U, v_k^* \otimes v_k > = p_k \Sigma \{ M^{jj'} N_k^{ii'} x_i^j \, x_{i'}^{j'} \mid j; j' \in J \; ; \; i, i' \in I \}$$

On reconnait l'expression de la mesure d'information associée au triplet (X^k, M, N_k) on a donc bien :

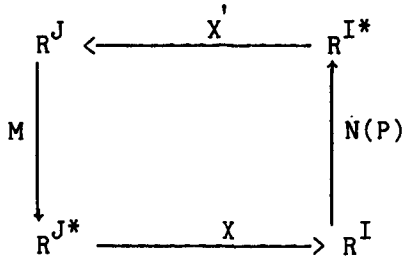
$$< U, v_k^* \otimes v_k > = p_k \, I(X^k, M, N_k)$$

En considérant alors :

le tableau $X = \pi \{X^k \mid k \in K\}$

l'espace $R^I = \sum \{R^{I_k} \mid k \in K\}$

la métrique $N(P) = \Sigma \{p_k N_k \mid k \in K\}$ diagonale par blocs, le schéma de dualité étant :



Les tenseurs Z, $Z_k \in R^{J*} \otimes R^J$ associés aux triplets $(X, M, N(P))$ et (X^k, M, N_k), on a les relations :

$$Z = \Sigma\{p_k Z_k \mid k \in K\} \text{ et } I(X, M, N(P)) = \Sigma\{p_k I(X^k, M, N_k) \mid k \in K\}$$

on a alors la proposition :

Proposition IV.2.5.1

On a les égalités :

$$I \text{ inter} = \Sigma \{< U, v_k^* \otimes v_k > \mid k \in K\} = I(X, M, N(P))$$

La métrique $N(P)$ étant telle que si $i \in I_k$, $i' \in I_{k'}$, si $k \neq k'$ $N_k^{ii'} = 0$ sinon $N_k^{ii'} = \frac{p_i}{p_k} \cdot \frac{p_{i'}}{p_k}$, la case (i, i') du tableau X est donc pondérée par les poids des individus i et i' relatif à leurs classes d'appartenance.

Nous avons une double décomposition des espaces E et F

$$E = \sum \{E_q \mid q \in Q\} \ , \ F = \sum \{F_k \mid k \in K\}$$

$$M = \Sigma \{c_q M_q \mid q \in Q\} \ , \ N(P) = \Sigma \{p_k N_k \mid k \in K\}$$

d'après le lemme III.4.2.1, on a :

$$\| X_E \otimes F \|^2_{M \otimes N(P)} = \Sigma \{c_q p_k \| X_{E_q} \otimes F_k \|^2_{M_q \otimes N_k} \mid q \in Q, k \in K\}$$

ou encore en termes d'information :

$$I(X, M, N(P)) = \Sigma \{c_q p_k I(X^k_q, M_q, N_k) \mid q \in Q, k \in K\}$$

Ainsi la classification automatique maximisant l'inertie inter maximise la somme pondérée des informations relatives aux classes p_k dans le contexte des métriques M_q, $q \in Q$ et N_k, $k \in K$.

IV.2.6 Choix des pondérations

Nous avons vu que pour une classe k donnée, l'inertie du nuage des individus dans l'espace total E est la moyenne de celle relative aux espaces E_q :

$$I_r (X^k, \underline{g}_k, m) = \Sigma \{c_q I^q_r (X^q_k, \underline{g}_{kq}, M_q) \mid q \in Q\}$$

On peut donc envisager les différentes possibilités examinées au paragraphe (III.5.6) que l'on rappelle :

1) Choisir des pondérations égales si les nuages d'individus par variables $N^I_{E_q}$ ont des inerties comparables sinon un nuage d'inertie importante influencera davantage dans la détermination des classes.

2) Normaliser les inerties des nuages $N^I_{E_q}$ c'est-à-dire choisir les coefficients c_q égaux à $1 \mid I(X_q, M_q, D_p)$ à un coefficient près.

3) Choisir les coefficients c_q qui maximise la liaison globale entre les variables $X_1, \ldots, X_Q$.

4) Lorsque l'on traite des questionnaires, il arrive souvent qu'une ou un ensemble de variables qualitatives jouent un rôle particulier ou bien l'étude d'un thème de l'enquête a permis de trouver une partition intéressante. On peut alors lorsque l'on étudie un ensemble de questions relatives à un autre thème de choisir (cf paragraphe (III.5.6)) une pondération maximisant la dépendance entre les questions et la partition estimée où la variable qualitative de référence.

Nous avons vu que cela revenait à pondérée chaque question par le degré de liaison : le phi-deux entre la question et la partition estimée ou variable qualitative de référence. Ainsi, on favorisera les questions les plus liées à la partition ou variable qualitative.

5) Le critère optimisé étant une somme de produits scalaires

$$W = \Sigma\{c_q \langle A_K, U_q \rangle \mid q \in Q\}$$

on peut choisir $c_q = 1/||A_K|| \; ||U_q||$ $q \in Q\}$ et W s'écrit :

$$W = \Sigma\{R_v(A_K, U_q) \mid q \in Q\}$$

En l'absence d'idée précise, on choisira cette solution ou la première.

IV.3 Analyse Factorielle Typologique Généralisée

IV.3.1 L'Analyse Factorielle Typologique [Ok 75] [Did 79]

Faisant remarquer que la contrainte d'orthogonalité des facteurs principaux ne permet pas la détection de tendances locales (directions d'allongement non orthogonales), ces auteurs proposent une approche plus générale de la simplification de la représentation euclidienne d'un nuage. Au lieu de chercher la droite affine, le plan, ou la variété affine de dimension q (q-variété affine) la plus proche du nuage à analyser, on recherche K-variétés affines de dimension q = 0, 1, ..., etc, les plus proches d'éventuels agglomérats locaux du nuage des individus N_J^I. Divers indices permettent d'évaluer la proximité entre un ensemble et variétés affines, entre le nuage initial et l'ensemble K-variétés affines.

a) Le modèle

Plus précisement on définit :

1) Une mesure de proximité entre un élément $\underline{x} \in R^J$ pour une métrique donnée M^* et H_r une variété affine de dimension r, $H_r \in \mathbb{H}_r$ qui désigne l'ensemble des variétés de dimension r de la manière suivante :

$$I_r : R^I \times \mathbb{H}_r \times M \longrightarrow R^+$$

$$\underline{x}, H_r, \times M^* \longrightarrow I_r(\underline{x}, H_r, M^*) = p(x) \inf \{d^2_{M^*}(\underline{x}, y) \mid y \in H_r\}$$

$$= p(x)\, d^2_{M^*}(\underline{x}, A_{H_r}(x))$$

où A_{H_r} désigne le projecteur orthogonal associé à H_r.

$I_r(\underline{x}, M^*, H_r)$ s'interpréte comme l'inertie de $\underline{x}$ par rapport à la variété H_r pour la métrique M^*

b) La mesure de proximité se prolonge aisément pour un ensemble $E' \subset R^J$ de la manière suivante :

On se donne une application M permettant d'associer à toute partie $E' \subset R^J$ une métrique $M(E') \in M$ ensemble des métriques définies sur R^J

$$M : \mathcal{P}(R^J) \longrightarrow \mathbf{M}$$

$$E' \longrightarrow M(E')$$

l'application I_r se définit sur $\mathcal{P}(R^J) \times H_r \times M$ comme suit :

$$I_r : \mathcal{P}(R^J) \times \mathbf{H} \times M \longrightarrow \mathbf{R}^+$$

$$(E', H_r, M(E')) \longrightarrow I_r(E', H_r, M(E')) = \Sigma \{I_r(\underline{x}, H_r, M(E')) \mid \underline{x} \in E'\}$$

$I_r(E', H_r, M(E'))$ s'interpréte comme l'inertie de E' relative à la variété affine H_r et mesure l'adéquation de H_r et de E'.

Définition :

L'Analyse Factorielle Typologique est la recherche d'une partition $P = \{P_k \mid k \in K\} \in \mathbb{P}_K$ ensemble des partitions à K classes, un ensemble K de r-variétés affines $H_r^K = \{H_r^k \mid k \in K\}$ résolvant le problème suivant :

Problème IV.3.1

$$\left|\begin{array}{l} \min \ \Sigma\{I_r(\underline{x}, H_r^k, M(P_k)) \mid \underline{x} \in P_k, \ k \in K\} \\ H_r^k \in H_r^K \\ P \in \mathbb{P}_K \end{array}\right.$$

critère exprimant l'adéquation entre les K-variétés affines et les K classes de P le problème IV.3.1 s'écrit alors :

Problème IV.3.2

$$\left|\begin{array}{l} \min \ \Sigma \{I_r (P_k, H_r^k, M(P_k)) \mid k \in K\} \\ H_r^k \in H_r^K \\ P \in \mathbb{P}_K \end{array}\right.$$

Pour K = 1 , le problème IV.3.2 s'écrit :

Problème IV.3.3

$$\left|\begin{array}{l} \min I_r(X, H_r, M(X)) \\ H_r \in H_r \end{array}\right.$$

On reconnait que la r-variété affine H_r solution de ce dernier problème est engendrée par les r-premiers axes factoriels de l'analyse du triplet $(X, M(X), D_p)$. L'Analyse Factorielle Typologique généralise les méthodes factorielles classiques.

IV.3.2 Généralisation à l'étude de tableaux n-aires

L'Analyse Factorielle Typologique se généralise facilement à Q triplets (X_q, M_q, D_p) en considérant les cadres de référence suivant :

- L'espace des individus $\mathbb{R}^J = \sum \{\mathbb{R}^{J_q} \mid q \in Q\}$

- Le tableau $X = \pi \{X_q \mid q \in Q\}$

- L'application M définie comme suit :

On se donne Q application M_q de $\mathcal{P}(\mathbb{R}^{J_q})$ dans $\mathbb{M}_q$ ensemble de métriques définies sur $\mathbb{R}^{J_q}$: $M_q : \mathcal{P}(\mathbb{R}^{J_q}) \longrightarrow \mathbb{M}_q$

$$E'_q \longrightarrow M_q(E'_q)$$

on considère alors l'espace produit $\mathcal{P}(\mathbb{R}^J) = \pi \{\mathcal{P}(\mathbb{R}^{J_q}) \mid q \in Q\}$ et l'application M tel que si $E' = \pi \{E'_q \mid q \in Q\} \in \mathcal{P}(\mathbb{R}^J)$, on définit M comme suit :

$$M : \mathcal{P}(\mathbb{R}^J) \longrightarrow \mathbb{M}$$

$$E' \longrightarrow M(E') = \Sigma \{c_q M_q(E'_q) \mid q \in Q\}$$

- le triplet considéré est (X, M, D_p) et l'AFTG est la recherche d'une partition $P = \{P_k \mid k \in K\} \in \mathbb{P}_k$ et un ensemble K de variétés affines $H_r^K = \{H_r^k \mid k \in K\}$ de $\mathbb{R}^J$ résolvant le problème suivant :

Problème IV.3.3

$$\left| \begin{array}{l} \min \ \Sigma \ I_r(P_k, H_r^k, M(P_k)) \mid k \in K \} \\ H_r^K \in H_r^K \\ P \in \mathbf{P}_K \end{array} \right.$$

avec $M(P_k) = \Sigma \{c_q M_q(P_k^q) \mid q \in Q\}$ où si $\underline{x} \in P_k$, $\pi_q \in P_k^q$.

IV.4 L'Analyse canonique Typologique Généralisée : AFTG

IV.4.1 Introduction

Dans l'introduction, nous avions attiré l'attention sur le fait qu'en général, la population d'individus I n'était pas homogène et qu'il était nécessaire de rechercher des liaisons locales entre les variables.

Le problème peut se formuler en les termes sivants : chercher une partition $P = (P_1, P_2, \ldots, P_K)$ de la population I tel que la somme des liaisons $\mathcal{L}(X_1^k, \ldots, X_Q^k)$ relatif à chaque classe P_k soit maximum. En terme de problème d'optimisation, cela s'écrit :

$$\max \ \Sigma \{\mathcal{L}(X_1^k, \ldots, X_Q^k) \mid k \in K\}$$
$$P \in \mathbf{P}_K$$

Nous avons défini $\mathcal{L}(X_1^k, \ldots, X_Q^k) = \max \Sigma \{c_q \mathcal{L}(v_k, X_q^k, M_q^k) \mid q \in Q\}$

$$\left| v_k \in R^{I_k} \text{ avec } \| v_k \|^2_{D_p} = 1 \right.$$

si l'on note a_k, le facteur sur R^J tel que $v_k = X_q^k a_k$, le problème d'optimisation se pose en les termes suivants :

Problème IV.4.1

$$\left|\begin{array}{l} \max W_1(a, P) = \Sigma \{c_q \mathcal{L}(X_q^k a_k, X_q^k, M_q^k) \mid q \in Q, k \in K\} \\ a \in R^{J \times K} \\ P \in \mathbf{P}^K \\ \text{avec } \| X_q^k a_k \|^2 D_p = 1, \; k \in K \end{array}\right.$$

Exemple :

Pour illustrer notre propos, considérons que nous avons un ensemble Q triplets (X_q, M_q, D_p) tels que les tableaux X_q soient des variables quantitatives centrées et $M_q = D_{1|\sigma^2}$, alors on a vu que :

$$\mathcal{L}(X_1, \ldots, X_Q) = \Sigma \{corr^2(\underline{x}^j, v) \mid j \in J\}$$

en accordant la même importance aux divers groupes $\{c_q = 1 \mid q \in Q\}$. La résolution du problème d'optimisation permettra donc de trouver une partition P de l'ensemble I et k composantes principales v_k, $v=(v_1,\ldots,v_k)$ tel que : $W(P, v) = \Sigma\{corr^2 (x_k^j, V_k) \mid j \in J, k \in K\}$, soit maximum.

les vecteurs v_k seront de bons "résumés" locaux de l'information et l'étude des variables $\underline{x}_k^j$ les plus liées avec v_k, permettra de détecter des associations locales $(corr^2 (\underline{x}_k^j, \underline{x}_k^{j'})$ important) entre variables.

IV.4.2 Lien entre l'AFTG et l'ACTG

Au paragraphe III.5.3, nous avons montré la symétrie qui existe dans les méthodes factorielles entre les problèmes d'optimisation de recherche de composantes principales et celle de recherche d'axes d'inertie minimum.

Cette dualité est importante car la résolution d'un problème permet d'avoir la solution à l'autre. Aussi, nous nous sommes intéressés aux liens entre les problèmes d'optimisations de recherches de k-composantes locaux (Analyse Canonique Typologique Généralisée) et celle de la recherche de k-axes d'inertie minimum (Analyse Factorielle Typologique Généralisée).

Précisons les problèmes :

Nous avons montré que :

$$\mathcal{L}(X_q^k\, a_k,\ X_q^K,\ M_q^k) = \langle X_{a_k}^k,\ W_q^k\, D_{p_k}\, X_{a_k}^k \rangle_{D_{p_k}}$$

par suite, le problème IV.4.1 s'écrit en notant $a = (a_1, \ldots, a_k)$ ou encore

Problème IV.4.1

$$\left|\begin{array}{l} \max W_1(a,p) = \Sigma\{c_q p_k \langle X_{a_k}^k, W_q^k\, D_{p_k}\, X_{a_k}^k \rangle \mid a \in Q,\ k \in K\} \\ a \in \mathbb{R}_K^{J\times K} \\ P \in \mathbb{P}^k \end{array}\right.$$

Problème IV.4.1.2

$$\left|\begin{array}{l} \max W_1(a,\ P) = \Sigma\ \{p_k \langle X_{a_k}^k,\ W_q^k\, D_{p_k}\, X_{a_k}^k \rangle_{D_{p_k}} \mid q \in Q,\ k \in K\} \\ a \in \mathbb{R}^{J\times K} \\ P \in \mathbb{P}_k \text{ sous les contraintes } \| v_k \|_{D_p}^2 = \| X^k a_k \|_{D_p}^2 = 1,\ k \in K \end{array}\right.$$

en se rappelant que $W^k\, D_{p_k} = \Sigma\ \{c_q\, W_q^k\, D_{p_k} \mid q \in Q\}$

nous pondérons les termes $\mathcal{L}(X_q^k\ a_k,\ X_q^k,\ M_q^k)$ par les poids des classes $p_k = \Sigma\ \{p(\underline{x}_i) \mid i \in I_k\}$, car les poids D_{p_k} sont normalisés à 1.

Soit $\mathbf{H}_1$, l'ensemble des variétés affines de dimension 1, le problème de recherche de k-vecteurs unitaires d'inertie minimum de l'Analyse Factorielle Typologique Généralisée en notant $u = \{u_k = H_1^k \mid k \in K\}$ est :

Problème IV.4.2.1

$$\left|\begin{array}{l} \min W_2(u, P) = \Sigma \{ I_r(P_k, u_k, M(P_k)) \mid k \in K \} \\ u \in \mathbb{R}^{J \times K} \\ P \in \mathbb{P}_K \text{ sous les contraintes } \| u_k \|^2_{M(P_k)} = 1 \text{ pour } k \in K \end{array}\right.$$

Il serait intéressant que la dualité observée dans l'étude de ces problèmes dans le cas de la population générale (partition à une classe) se conserve dans l'étude de K classes. En effet, il suffit de résoudre un problème pour avoir la solution de l'autre et aux k-axes d'inertie minimum correspondraient k-composantes principales optimales non seulement au niveau local de la classe mais, aussi, du point de vue global au niveau des partitions P de $\mathbb{P}_K$.

Nous allons montrer que c'est le cas pour les métriques usuelles (Analyse Canonique Généralisée, Analyse en Composantes Principales, Analyse des Correspondances Multiples).

Proposition IV.4.2.1

Si pour chaque groupe de variables X_q, on choisit l'une des métriques suivantes :

$M_q^k = (V_{qq}^k)^{-1}$ métrique de Mahalanobis

$M_q^k = D_{1|\sigma_k^2}$ métrique de Sebestyen

$M_q^k = D_{|P_q^k}$ métrique du chi-deux pour une variable qualitative

alors l'Analyse Canonique Typologique Généralisée et l'Analyse Factorielle Typologique Généralisée sont équivalentes.

Démonstration :

En d'autres termes les problèmes (IV.4.1). et (IV.4.2.1) sont équivalents. La résolution de (IV.4.2.1) se fait en deux temps, pour une partition $\mathbb{P} \in \mathbb{P}_K$ fixée,

on cherche $u \in \mathbb{R}^{J \times K}$ solution de :

Problème IV.4.2.2

$$\left| \begin{array}{l} \min W_2(u, P) = \Sigma \{ I_r(P_k, u_k, M^k) \mid k \in K \} \\ u \in \mathbb{R}^{J \times K} \end{array} \right.$$

Les Analyses en Composantes Principales des K-triplets (X^k, M_k, D_{pk}) montrent que u_k est le vecteur unitaire de l'axe factoriel associé à la plus grande valeur propre λ_k et :

$$I_r(P_k, u_k, M^k) = P_k . (\text{trace } V_k M^k - \lambda_k)$$

La présence du facteur p_k est due au fait que nous n'avons pas normalisé les poids dans la définition de $I_r(P_k, u_k, M^k)$

Par suite le minimum est :

$$\left| \begin{array}{l} \min W_2(u, P) = \Sigma \{ p_k \text{ trace } V_k M^k - \lambda_k p_k \mid k \in K \} \\ u \in R^{J \times K} \end{array} \right.$$

rappelons que :

$$V_k M^k = \begin{bmatrix} V_{11}^k & \cdots & V_{1Q}^k \\ \vdots & & \vdots \\ V_{Q1}^k & \cdots & V_{QQ}^k \end{bmatrix} \begin{bmatrix} c_1 M_1^k & & \\ & \ddots & \\ & & c_Q M_Q^k \end{bmatrix}$$

d'où :

$$\text{trace } V_k M^k = \Sigma \{ c_q \text{ trace } V_{qq}^k M_q^k \mid q \in Q \}$$

On peut partager l'ensemble des groupes de variables en 2 groupes Q_1 et Q_2, $Q = Q_1 \quad Q_2$ où Q_1 est l'ensemble des groupes de variables X_{q_1} centrées de type quantitatif et Q_2 l'ensemble des groupes de variables X_{q_2} de type qualitatif centrée.

Par hypothèse : $\forall\, q_1 \in Q_1 \quad M^k_{q_1} = (V^k_{q_1 q_1})^{-1}$ ou $M^k_{q_1} = D_{1/\sigma^2_k}$

par suite : trace $V^k_{q_1 q_1} M^k_{q_1} = M^k_{q_1} = \text{card } J_{q_1}$ et l'on a :

$$\Sigma \{c_{q_1} \text{ trace } V^k_{q_1 q_1} M^k_{q_1} \mid q_1 \in Q_1\} = \Sigma \{c_{q_1} \text{ card } J_{q_1} \mid q_1 \in Q_1\}$$

Pour l'ensemble des variables qualitatives $\{X_{q_2} \mid q_2 \in Q_2\}$, $M^k_{q_2} = 1/P^k_{q_2}$ d'où trace $V^k_{q_2 q_2} M^k_{q_2} = \text{card } J_{q_2}$, on a donc aussi :

$\Sigma \{c_{q_2} \text{ trace } V^k_{q_2 q_2} M^k_{q_2} \mid q_2 \in Q_2\} = \Sigma \{c_{q_2} \text{ card } J_{q_2} \mid q_2 \in Q_2\}$ et alors trace $V_k M^k = \Sigma \{c_q \text{ card } J_q \mid q \in Q\} = C$ une constante alors :

$$\Sigma \{p_k \text{ trace } V_k M^k \mid k \in K\} = C \times \Sigma \{p_k \mid k \in K\} = C$$

d'où le problème d'optimisation (IV.4.2.3) s'écrit :

Problème IV.4.2.3

$$\begin{cases} \min \quad W_2(u, P) = C - \Sigma \{p_k \lambda_k \mid k \in K\} \\ u \in R^{J \times K} \\ \text{avec } \| u_k \|^2_{M^k} = 1, \; k \in K \end{cases}$$

Nous avons vu au paragraphe que la liaison $(X^k_1, \ldots, X^k_Q)$ était égale à la première valeur propre λ_k de l'ACP de (X^k, M^k, D_{p_k})

$$\lambda_k = (X^k_1, \ldots, X^k_Q) = \max \left\{ < X^k_{a_k}, W^k D_{p_k} X^k_{a_k} >_{D_{p_k}} \;\middle|\; \begin{array}{l} a \in R^J \\ \text{avec } \| X^k_{a_k} \|^2_{D_{p_k}} = 1, \; k \in K \end{array} \right.$$

On a donc :

$$\Sigma \{p_k \lambda_k \mid k \in K\} = \max W_1(a, P) \Big| \begin{array}{l} a \in \mathbb{R}^{J \times K} \\ \text{avec } \| X^k_{a_k} \|^2_{D_{p_k}} = 1, \ k \in K \end{array}$$

nous avons donc la relation entre les deux problèmes :

$$\min W_2(u, P) = C - \max W_1(a, P) \Big| \begin{array}{l} a \in \mathbb{R}^{J \times K} \\ \text{avec } \| X^k_{a_k} \|^2_{D_{p_k}} = 1, \ k \in K \end{array}$$
$$\Big| \begin{array}{l} u \in \mathbb{R}^{J \times K} \\ \text{avec } \| u_k \|^2_{M^k} = 1, \ k \in K \end{array}$$

Nous en déduisons donc l'équivalence entre les deux problèmes d'optimisation :

$$\min W_2 (u, P) = C - \max W_1 (a, P)$$
$$\Big| \begin{array}{l} u \in \mathbb{R}^{J \times K} \\ \text{tel que } \| u_k \|^2_{M_h} = 1, \ k \in K \\ P \in \mathbb{P}_K \end{array} \qquad \Big| \begin{array}{l} a \in \mathbb{R}^{J \times K} \\ \text{sous les conditions } \| X^k a_k \|^2_{D_{p_h}} = 1 \\ P \in \mathbb{P}_K \end{array}$$

D'une manière générale, pour les métriques $M \in \mathbb{M}$ tel que

$$\forall E \subset \mathbb{R}^I, \quad I_r (E, G, M(E)) = C \text{ constante}$$

alors, les deux problèmes d'optimisation sont équivalents. Les deux critères étant liés par la relation :

$$\min W_2 (u, P) = K \times C - \max W_1 (a, P)$$
$$\Big| \begin{array}{l} u \in \mathbb{R}^{J \times K} \\ \text{tel que } \| u_k \|^2_{M_k} = 1, \ k \in K \\ P \in \mathbb{P}_K \end{array} \qquad \Big| \begin{array}{l} a \in \mathbb{R}^{J \times K} \\ \text{tel que } \| X^k a_k \|^2_{D_{p_k}} = 1 \\ P \in \mathbb{P}_K \end{array}$$

en effet, trace $(V_k M^k) = I (P_k, G_k, M^k) = C$

Etude du cas où $M^k = M$ pour $k \in K$

Dans cette partie, nous étudions le cas où la métrique est fixée quelque soit la classe P_k de la partition P .

Ainsi, $M_q^k = M_q \quad \forall k \in K$ et $\forall q \in Q$

Ces métriques étant définies par exemple relatives à toute la population I. Rappelons la relation classique :

$$I_r((P_k, u_k, M) + I_r (P_k, u_k^{\perp}, M) = I_r (P_k, G_k, M) \qquad (IV.4.1)$$

entre l'inertie de la classe P_k et l'inertie relative à u_k et le vecteur M-orthogonal $u_k^{\perp}$, G_k étant le centre de gravité de P_k.

Considérons le problème d'optimisation suivant :

Problème IV.4.2.4

$$\begin{cases} \max W_3 (u, P) = \Sigma \{I_r (P_k, u_k^{\perp}, M) \mid k \in K\} \\ u \in \mathbb{R}^{J \times K} \\ \text{avec } \| u_k \|^2_M = 1 \\ P \in \mathbb{P}_K \end{cases}$$

Le problème (IV.4.2.4) est équivalent à (IV.4.1) relatif à l'ACTG, car, pour une partition P :

$$\begin{cases} \max W_3 (u, P) = \Sigma \{p_k \lambda_k \mid k \in K\} \\ u \in \mathbb{R}^{J \times K} \\ \text{avec } \| u_k \|^2_M = 1 \\ P \in \mathbb{P}_K \end{cases}$$

Etudions les liaisons entre les critères W_3 et W_2. A partir de l'expression (IV.4.1), on a :

$$\Sigma\{I_r(P_k, u_k, M) \mid k \in K\} + \Sigma\{I_r(P_k, u_k^{\perp}, M) \mid k \in K\} = \Sigma\{I_r(P_k, G_k, M) \mid k \in K\}$$

comme $\Sigma\{I_r(P_k, G_k, M) \mid k \in K\} = I_r(P, G, M) - \Sigma\{p_k\, d_M^2(G_k, G) \mid k \in K\}$

d'après la relation classique entre :

l'inertie inter : $Iinter = \Sigma\{p_k\, g^2(G_k, G) \mid k \in K\}$

et l'inertie intra : $Iintra = \Sigma\{I_r(P_k, G_k, M) \mid k \in K\}$

d'où :

$$\Sigma\{I_r(P_k, u_k, M) \mid k \in K\} + \Sigma\{p_k d^2(G_k, G) \mid k \in K\} + \Sigma\{I_r(P_k, u_k^{\perp}, M) \mid k \in K\} = I(P, G, M)$$

ou encore :

$$W_2(u, P) + W_3(u, P) + Iinter = I_r(P, G, M) = \text{constante}$$

Il apparaît que, si l'on maximise uniquement $W_3(u, P)$, on minimise $W_2(u, P) + Iinter$. On risque alors d'obtenir des classes peu écartées du centre de gravité général G (Inertie inter petite). Ce qui est intéressant, c'est de maximiser $W_3(u, P) + Iinter$ ou minimiser $W_2(u, P)$, c'est-à-dire faire une Analyse Factorielle Typologique Généralisée. Priviligier la recherche d'axes d'inerties minimum.

En résumé, lorsque la métrique est indépendante des classes, le problème d'optimisation que nous résolvons est :

Problème IV.4.2.5

$$\left|\begin{array}{l} \max \Sigma\{p_k < X a_k^k,\ W^k D_{p_k} X a_k^k >_{D_{p_k}} \mid k \in K\} + \Sigma\{p_k d_M^2(G_k, G) \mid k \in K\} \\ a \in \mathbb{R}^{J \times K} \\ \text{avec } \| X^k a_k \|^2_{D_{p_k}} = 1 \\ P \in \mathcal{P}_K \end{array}\right.$$

IV.4.3 Interprétation des critères

Les critères optimisés par la méthode proposée s'interprète différemment suivant l'espace considéré, individu ou variable.

Dans l'espace des individus, la méthode est une méthode de type factorielle c'est-à-dire cherchant des axes factoriels locaux d'inertie minimum permettant donc de détecter des agglomérats locaux.

Dans l'espace des variables, la situation est plus complexe étant donné la nature diverse des variables, les métriques pouvant, non seulement, varier suivant les types de variables, mais aussi, selon les classes suivant l'algorithme choisi (métrique globale ou locale).

Dans un premier temps, nous donnons une interprétation de l'algorithme en terme général de décomposition d'information et préciserons, dans les applications, la nature exacte des liaisons suivant les types de tableaux et de métriques choisis.

1er cas : les métriques M_q^k sont telles que $I_r(X_q^k, M_q^k, D_{p_k}) = c_q$.

Ce sont les métriques de la proposition tel que l'inertie d'une partie de E soit constante quelque soit E. Soit U_q^k, l'opérateur associé au triplet (X_q^k, M_q^k, D_{p_k}), l'opérateur associé à (X^k, M^k, D_{p_k}) est noté U^k :

$$U^k = \Sigma \{c_q U_q^k \mid q \in Q\}$$

au niveau des informations associées aux triplets (X_q^k, M_q^k, D_{p_k}), on a la relation :

$$I(X^k) = \Sigma \{c_q I(X_q^k) \mid q \in Q\} = \Sigma \{c_q c_q \mid q \in Q\} = \text{constante}$$

On a la décomposition de $I(X^k)$ en fonction de $u_k \in R^J$, $v_k \in R^{I_k}$ est :

$$I(X^k) = < Z_k, u_k^* \otimes u_k > + < Z_k, \overset{\perp M_k}{\otimes} u_k >$$

$$= < U_k, v_k^* \otimes v_k > + < U_k, \overset{\perp M_k}{\otimes} v_k >$$

L'information relative au tableau total X est puisque $I(X^K) = C$

$$I(X^K) = \Sigma \{p_k\ I(X^K) \mid k \in K\} = C$$

L'Analyse canonique généralisée typologique maximise le critère W_1 s'exprimant :

$$W_1 = \Sigma \{p_k < U_k, v_k^* \otimes v_k > \mid k \in K\}$$

L'Analyse factorielle typologique généralisée minimise le critère W_2 :

$$W_2 = \Sigma \{p_k < Z_k, \otimes^{\perp M_k} u_k > \mid k \in K\}$$

Le critère W_3 s'écrit :

$$W_3 = \Sigma \{p_k < Z_k, u_k^* \otimes u_k \mid k \in K\}$$

Les études précédentes montrent que les problèmes de maximisation de W_1 et W_3 sont identiques par suite :

$$C = W_1 + W_2$$

L'Analyse factorielle typologique généralisée et l'Analyse canonique généralisée typologique sont donc identiques, car minimiser W_2 revient à maximiser W_1.

Ainsi donc, ces méthodes s'interprétent comme la recherche d'une décomposition optimale de l'information I(X) par un ensemble de k-opérateurs U'_k et Z'_k et k vecteurs v_k et u_k relatifs à une partition P à k classes minimisant e :

$$e = I(X) - \Sigma \{ < U'_k, v_k^* \otimes v_k > \mid k \in K\}$$

et

$$e = I(X) - \Sigma \{ < Z'_k, u_k^* \otimes u_k > \mid k \in K\}$$

$\{U'_k \mid k \in K\}$ étant l'opérateur $U'_k = p_k\ U_k$ du triplet (X^K, M^K, D_{p_k}) relatif à la classe P_k de poids p_k.

$\{Z'_k \mid k \in K\}$ étant l'opérateur $Z'_k = p_k\ Z_k$ du triplet (X^K, D_{p_k}, M^K).

2^{e} cas : les métriques M_q^k sont constantes et égales à M_q.

Nous partirons de la décomposition.

$$Ir(P,G,M) = Iinter + Iintra$$

nous supposerons que le tableau X est centré, on a alors la mesure de l'information $I(X,M,D_p) = I_r(P,G,M)$. Nous avons vu que

$I_{Inter} = \Sigma\{<U, v_k^* \otimes v_k> | k \quad K\}$ où v_k sont les variables indicatrices de la variable qualitative X_k associée à la partition P. On a

$I_{Intra} = \Sigma\{Ir(P_k,G_k,M\}$ et si on note $\tilde{X}_k$ le tableau centré relatif à P_k, nous avons le lien entre l'inertie et la mesure de l'information relative à la classe P_k :

$$Ir(P_k,G_k,M) = I(\tilde{X}_k,M,D_{pk}) = <Z_k,u_k^* \otimes u_k> + <Z_k,\overset{\perp M}{\otimes} u_k> \text{ où}$$

$Z_k \in E^* \otimes E$ le tenseur associé à $(\tilde{I}_k,M,D_{pk})$ d'où

$$I(X,M,D_p) = \Sigma\{<U,v_k^* \otimes v_k> | k \in K\} + \Sigma\{<Z_k,u_k^* \otimes u_k> | k \in K\} + \Sigma\{<Z_k,\overset{\perp M}{\otimes} u_k> | k \quad K\}$$

la méthode proposée recherche donc une décomposition optimale de $I(X,M,D_p)$ à l'aide des vecteurs $\{v_k | k \in K\}$ de E et $\{u_k | k \in K\}$ de E et on ne peut rien dire de plus car $X + \pi\{\tilde{X}_k | k \in K\}$ les données étant centrées localement par classe.

IV.4.4 Construction de l'algorithme

L'algorithme construit est de type "Nuées Dynamiques" dans le cas de métriques globales et ne pose pas de difficultés. Lorsque les métriques sont locales, on est obligé d'avoir un algorithme de type transfert. On note :

$\mathbf{L} = H_1 \times \mathbf{M}$ l'espace dont les éléments sont des couples constitués d'un sous-espace affine de dimension 1 : $u \in H_1$ et d'une métrique $M \in \mathbf{M}$.

$\mathbf{L} = (u, M) \in \mathbb{L}$

et $\mathbf{L}^K = (H_1 \times \mathbf{M})^K$ l'espace produit si $L \in \mathbf{L}^K$, alors :

$L = (L_1, \ldots, L_K)$ où $L_j = (u_j, M_j) \in \mathbf{L}$

On considère une fonction g de $\mathcal{P}(R^J)$ dans $\mathbf{L}$ dit fonction de représentation qui est la composée de deux fonctions 1 et h définies de la manière suivante :

$$\begin{array}{ccccc} \mathcal{P}(R^J) & \xrightarrow{\quad 1 \quad} & \mathcal{P}(R^J) \times M & \xrightarrow{\quad h \quad} & \mathbf{L} \\ E & \longrightarrow & (E, M(E)) & \longrightarrow & (u, M(E)) \end{array}$$

où u est le vecteur directeur de l'axe principal d'inertie de $E \subset R^J$ muni de la métrique M(E). Ainsi :

$$\begin{array}{ccc} \mathcal{P}(\mathbb{R}^J) & \xrightarrow{\quad g \quad} & \mathbb{L} \\ E & \longrightarrow & g(E) = h \circ 1(E) = (u, M(E)) \end{array}$$

et g_K, l'application de $\mathbf{P}_K$ dans $\mathbf{L}^K$ définie comme suit :

$$\begin{array}{ccc} \mathbf{P}_K & \xrightarrow{\quad g_K \quad} & \mathbf{L}^K \\ P & \longrightarrow & g_K(P) = (g(P_1), \ldots, g(P_K)) \end{array}$$

permet de déterminer le K-centre d'aggrégation $L = (L_1, \ldots L_K)$.

On définit, ensuite, la fonction d'affectation f qui, à partir d'un K-centre d'aggrégation, détermine la partition $P \in \mathbb{P}_K$ la plus adaptée :

$$\begin{array}{ccc} \mathbf{L}^K & \xrightarrow{\quad f \quad} & \mathbf{P}_K \\ L & \longrightarrow & f(L) = P = (P_1, \ldots, P_K) \end{array}$$

Proposition IV.4.4.1

Soit le critère $W_2(L, P) = \Sigma \{I(P_k, u_k, M(P_k)) \mid k \in K\}$, il est possible par l'intermédiaire des fonctions g et f judicieusement choisies de construire une suite (P^n, L^n) faisant converger la suite $u_n = W(P^n, L^n)$.

1er cas : la métrique $M(P_k) = M$ est constante, indépendante de la classe P_k.

La fonction g : P_k dans L^k étant définie comme précédemment, précisons la fonction f.

Soit $L = (L_1, \ldots, L_K) \in L^K$, définissons la distance d'un individu $\underline{x}$ à $L_j = (U_j, M) = p(\underline{x})\, d^2_M [\underline{x}, A_j(\underline{x})]$, alors $P = (P_1, \ldots, P_K)$ est définie comme suit :

$$P_i = \{\underline{x} \in E \mid D(\underline{x}, L_i) \leq D(\underline{x}, L_j) \text{ et } i < j \text{ en cas d'égalité}\}$$

La suite $u_n = W(P^n, L^n)$ est décroissante, en effet :

$$u_n = W(P^n, L^n) \geq W(P^n, L^{n+1}) \geq W(P^{n+1}, L^{n+1}) = u_{n+1}$$

avec $L^{n+1} = g(L^n)$ et $P^{n+1} = f(P_n)$. La première inégalité est vraie à cause de la définition de g (L^{n+1} est l'axe d'inertie minimum). La deuxième inégalité est vraie car :

$$\Sigma \{I(P^n_k, u^n_k, M) = \Sigma \{I(\underline{x}, u^n_k, M) \mid \underline{x} \in P^n_k, k \in K\}$$

$$= \Sigma \{D(\underline{x}, L_k) \mid \underline{x} \in P^n_k, k \in K\}$$

$$\geq \Sigma \{D(\underline{x}, L_k) \mid \underline{x} \in P^{n+1}_k, k \in K\}$$

En effet, par définition de f, tout élément $\underline{x}$ est plus proche de sa classe, dans la partition P^{n+1} que dans la partition P^n. La suite u_n, étant décroissante et minorée par 0, converge.

<u>2ème cas</u> : la métrique $M(P_k)$ n'est pas constante sur les classes, alors, la fonction f doit être modifiée.

On définit une fonction variation ΔI exprimant la variation d'inertie de $E \subset \mathbb{R}^I$ relativement à un espace affine $H_r \in \mathcal{H}_r$ à cause de $\underline{z} \in \mathbb{R}^J$ de la manière suivante :

$$\mathcal{P}(\mathbb{R}^J) \times H_r \times M \times \mathbb{R}^J \xrightarrow{\quad \Delta I \quad} \mathbb{R}^+$$

$$(E, H_r, M, \underline{z}) \longrightarrow \Delta I\,(E, H_r, M, \underline{z})$$

tel que :

$$\underline{z} \notin E \qquad \Delta I\ (E, H_r, M, \underline{z}) = I\big(E, H_r, M(E + \{\underline{z}\})\big) - I\big(E, H_r, M(E)\big)$$

$$\underline{z} \in E \qquad \Delta I\ (E, H_r, M, \underline{z}) = I\big(E, H_r, M(E - \{\underline{z}\})\big) - I\big(E, H_r, M(E)\big)$$

La distance d'un élément $\underline{z}$ à un représentant $L_k\ (u_k, M_k)$ s'écrit :

$$D(\underline{z}, L_k) = I(\underline{z}, u_k, M_k) + \Delta I\ (P_k, u_k, M_k, \underline{z}) + \Delta I\ (\underline{z}, u_k, M_k, \underline{z})$$

Dans la suite, on notera :

$$D(\underline{z}, L_k) = D(\underline{z}, u_k, P_k) \text{ car les classes } P_k \text{ vont évoluées } u_k \text{ restant fixe.}$$

Lorsque la métrique M est constante, $\Delta I = 0$ et on retrouve la distance précédente.

On définira la fonction d'affectation que l'on notera F, l'application de $\mathbb{P}_K \times \mathcal{P}(\mathbb{R}^J) \times H_1^K$ dans $\mathbb{P}_K$ qui à une partition $Q \in P_K$, une partie E de $\mathbb{R}^I$ et K-uple de variétés linéaires $u = (u_1, \ldots, u_k)$ associe une nouvelle partition P de k classes maximum. La fonction F se construit comme dans les nuées dynamiques séquentielles [Did75], l'idée étant de faire changer un élément que s'il améliore le critère.

$$\mathbb{P}_K \times \mathcal{P}(\mathbb{R}^J) \times H_1^K \longrightarrow \mathbb{P}_K$$

$$(Q, E, u) \longrightarrow F(Q, E, u) = P$$ où P est définie à l'aide d'une suite $\{\pi_n\}$ de $\mathbb{P}_K$ comme suit : on note

$\pi_\ell = (\pi_{\ell,1}, \ldots, \pi_{\ell,k})$ et $E = \{\underline{z}_1, \ldots, \underline{z}_r\}$ où $z_i \in \mathbb{R}^I$, r est donc le nombre d'éléments de E.

Soit $\pi_0 = Q$, la suite se définit par récurrence à partir de π_0 : étant donné la partition $\pi_{\ell-1}$ et $\underline{z}_\ell$, un élément de la classe j de cette partition $z_\ell \in \pi_{\ell-1,j}$ on construit la partition π_ℓ comme suit :

a) $i : D(z_\ell, u_i, \pi_{\ell-1,i}) \leq D(z_\ell, u_j, \pi_{\ell-1,j})$ avec $i<j$ en cas d'égalité alors $\pi_\ell = \pi_{\ell-1}$

b) $i : D(z_\ell, u_i, \pi_{\ell-1,i}) = \underset{r}{\text{Min}}\, D(z_\ell, u_r, \pi_{\ell-1,r})$

$$< D(z_\ell, u_j, \pi_{\ell-1,j})$$

avec $i<j$ en cas d'égalité, alors z_ℓ est affecté à la classe i de π_ℓ

$$\pi_{\ell,i} = \pi_{\ell-1,i} \cup \{z_\ell\}$$

et

$$\pi_{\ell,j} = \pi_{\ell-1,j} - \{z_\ell\}$$

les autres classes restant inchangées : $\pi_{\ell,p} = \pi_{\ell-1,p}$ $\forall$ p $\in$ K tel que $p \neq i$ et $p \neq j$.

Par récurrence, on calcule $\pi_1, \pi_2, \ldots, \pi_r$ et on pose $P = \pi_r$.

Montrons que $W(P^n, u^n, M(P^n)) \geq W(P^{n+1}, u^{n+1}, M(P^{n+1}))$.

Deux cas sont à considérer, soit aucun individu z ne change de classe, alors $P^n = P^{n+1}$, soit un individu a change de classe , il existe donc une étape ℓ et deux indices i et j tel que :

$$D(z_\ell, u_i^n, \pi_{\ell-1,\,i}) \leq D(z_\ell, u_j^n, \pi_{\ell-1,\,j})$$

Calculons la variation du critère à cette étape, seul les termes relatifs à la classe i et j, classe de Z_ℓ dans $\pi_{\ell-1}$, de la partition $\pi_{\ell-1}$ ont changé.

$$\Delta W = W(\pi_\ell, u^n, M_K(\pi_\ell)) - W(\pi_{\ell-1}, u^n, M_K(\pi_{\ell-1})) = \Delta_i - \Delta_j \text{ où}$$

$$\Delta_i = I(\pi_{\ell,i}, u_i^n, M(\pi_{\ell,i})) - I(\pi_{\ell-1,i}, u_i^n, M(\pi_{\ell-1,i}))$$

$$\Delta_j = I(\pi_{\ell,j}, u_j^n, M(\pi_{\ell,j})) - I(\pi_{\ell-1,j}, u_j^n, M(\pi_{\ell-1,j}))$$

comme : $\pi_{\ell,i} = \pi_{\ell-1,i} + \{z_\ell\}$ et $\pi_{\ell,j} = \pi_{\ell-1,j} - \{z_\ell\}$

par suite :

$$\Delta_i = I(\pi_{\ell-1,i}, u_i^n, M(\pi_{\ell,i})) + I(z_\ell, u_i^n, M(\pi_{\ell,i})) - I(\pi_{\ell-1,i}, u_i^n, M(\pi_{\ell-1,i}))$$

$$= I(z_\ell, u_i^n, M(\pi_{\ell-1,i})) + \Delta I(\pi_{\ell-1,i}, u_i^n, M(\pi_{\ell-1,i}), z_\ell)$$

$$+ \Delta I(z_\ell, u_i^n, M(\pi_{\ell-1,i}), z_\ell)$$

$$= D(z_\ell, u_i^n, \pi_{\ell-1,i})$$

symétriquement : $\Delta_j = D(z_\ell, u_j^n, \pi_{\ell-1,j})$, par suite :

$$\Delta W = D(z_\ell, u_i^n, \pi_{\ell-1,i}) - D(z_\ell, u_j^n, \pi_{\ell-1,j}) < 0$$

CQFD.

IV.5 Deux méthodes particulières

Nous explicitons dans ce paragraphe, les critères optimisés par l'AFTG, dans deux cas particuliers importants correspondant à l'analyse en composantes principales et l'analyse des correspondances multiples. Nous nous réservons, dans un prochain article, d'exposer la mise en oeuvre pratique, les aides à l'interprétation, l'étude de la stabilité des résultats et l'application à des données réelles.

IV.6 L'Analyse en Composantes Principales Typologiques

Cette méthode a été étudiée, dans le cas d'un tableau de mesures par OK [OK75], le point de vue étant, alors, celui des méthodes factorielles c'est-à-dire l'étude du nuage des individus par la recherche de K-variétés affines de dimension r d'inertie minimum. D'une manière générale, comme il a été déjà dit, la métrique sur les variables D_p étant toujours diagonale, AFTC s'interprétera toujours de cette façon si l'on se place du point de vue des individus. Nous

allons expliciter les critères optimisés en examinant l'ensemble des variables. On considère ici que l'on étudie un ensemble de triplets de tableaux quantitatifs (X_q, M_q, D_p), $q \in Q$, on distinguera deux cas.

1) Métriques locales : $M_q^k = D_{1|\sigma_k^2}$, $k \in K$, $q \in Q$

c'est-à-dire pour chaque ensemble X_q^k, on choisit la métrique de Sebestien adoptée à la classe. Chaque analyse de la classe k est donc une ACP normée pondérée par les coefficients c_q; Le critère optimisé est alors (cf IV.4) :

$$W(v, P) = \Sigma \{\mathcal{L}(X_1^k, \ldots, X_Q^k) \mid k \in K\}$$

$$= \Sigma \{c_q \ \text{corr}^2(\underline{x}_k^j, v_k) \mid j \in J_q, \ q \in Q, \ K \in K\}$$

(IV.5.1)

où $v = \pi \{v_k \mid k \in K\}$ tel que $\| v_k \|_{D_p}^2 = 1$, $k \in K$.

2) Métrique globale : $M_q^k = D_{1|\sigma^2}$, $q \in Q$, $k \in K$:

Nous avons vu alors que le critère optimisé est (cf problème IV.4.2.5) :

$$W(v, P) = \text{Inter} + \Sigma \{\mathcal{L}(X_1^k, \ldots, X_Q^k) \mid k \in K\}$$

calculons : $\mathcal{L}(X_1^k, \ldots, X_Q^k)$

$$\mathcal{L}(X_1^k, \ldots, X_Q^k) = \Sigma \left\{ \frac{c_q}{\text{var}\underline{x}^j} \text{covar}^2(\underline{x}_k^j, v_k) \mid j \in J_q, \ q \in Q, \ k \in K \right\}$$

(IV.5.2)

$$\mathcal{L}(X_1^k, \ldots, X_Q^k) = \Sigma \left\{ c_q \ \text{corr}^2(\underline{x}_k^j, v_k) \ \frac{\text{var}\underline{x}_k^j}{\text{var}\underline{x}^j} \mid j \in J_q, \ q \in Q, \ k \in K \right\}$$

En comparant les expressions (IV.5.1) et (IV.5.2), on voit que le choix des métriques locales accorde la même importance aux variables, tandis que, dans le cas de métriques globales, les variables ayant une forte variance dans la population totale interviendront peu dans l'analyse.

IV.6 L'Analyse des Correspondances Multiples Typologiques

Deux cas sont aussi à considérer, suivant que la métrique est adaptée ou non à la classe.

Nous avons donc un ensemble de triplets $(X_q, D_{1|P_q}, D_p)$, $q \in Q$, relatifs à des variables qualitatives. Les résultats du paragraphe précédent se transposent facilement en remarquant que si les variables qualitatives ne sont pas centrées au lieu de parler de "variance", on utilisera la "norme" pour la métrique D_p et au lieu de la "corrélation", on parlera de "cosinus".

Ainsi, si les métriques sont adaptées aux classes, c'est-à-dire, on effectue réellement une correspondance multiple pondérée sur les classes, le critère optimisé est :

$$W(v, P) = \Sigma \{c_q\, p_k \cos^2(\underline{x}_k^j, v_k) \mid j \in J_q,\ q \in Q,\ k \in K\}$$

où $\underline{x}_k^j$ est la modalité j de la variable q

Remarquons que : $\|\underline{x}_k^j\|^2_{D_p} = \Sigma \{p_i\, x_i^{j\,2} \mid i \in I_k\}$

$$= \frac{n_j^k}{n}$$

si n_j^k est l'effectif de la modalité j dans la classe k. Par suite, lorsque la métrique est globale, le critère optimisé est :

$$\mathcal{I}(X_1, \ldots, X_Q) = \Sigma \{c_q\, p_k \cos^2(\underline{x}_k^j, v_k) \frac{n_j^k}{n} \mid j \in J_q,\ q \in Q,\ k \in K\}$$

ainsi, les modalités d'effectif important dans la population totale interviendront peu dans une telle analyse.

V CONCLUSIONS

Nous avons défini la mesure d'information d'un tableau comme la norme du tenseur associé et étudié la réduction de cette information. Cette démarche semble être fructueuse car elle nous a permis de généraliser les approches de Caroll et d'Escofier pour les méthodes factorielles et les méthodes de classification type

Nuées Dynamiques.

Il s'agit maintenant d'approfondir la mise en oeuvre pratique des méthodes d'Analyse en Composantes Principales et Correspondances Multiples typologiques.

N.B. Je remercie le Professeur Cazes de l'Université Paris dauphine pour les diverses corrections et remarques qu'il a apporté à cette première partie.

BIBLIOGRAPHIE

<Ben 73> BENZECRI J.P. & al
L'Analyse des Données. Tome 1, La taxinomie, Tome 2, l'Analyse des Correspondances. Dunod. Paris, 1973.

<Ben 82> BENZECRI J.P.
Sur la généralisation d'un tableau de Burt et son analyse par bandes. Les cahiers de l'Analyse des Données, VOL I, n°1, 1982 pp 33-43

<Bra 73> BRAUN J.M.
Contribution à l'étude des séries chronologiques multiplies par l'Analyse des Données. Thèse de 3ème cycle, Université de Paris VI, 1973.

<Cac 70> CARROL J.D. et CHANG J.J
Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition Psychometrika, VOL 35, n°3, pp 283-319

<CaP 76> CAILLEZ F. et PAGES J.P.
Introduction à l'Analyse des Données. SMASH, 1976

<Caz 80> CAZES P.
L'analyse de certains tableaux rectangulaires décomposés en blocs. Les cahiers de l'Analyse des Données, VOL 4, n°4, 1980, pp 387-406.

<Did 75> DIDAY E.
Classification automatique séquentielle pour grands tableaux RAIRO Intelligence Artificielle et Reconnaissance des formes, Mars 1975.

<Did 78> DIDAY E.
Analyse canonique du point de vue de la classification automatique. Rapport INRIA. Rocquencourt 1978

<Did 79> DIDAY E. & Collaborateurs
Optimisation en Classification Automatique. INRIA Rocquencourt 1979.

<EsP 82> ESCOFIER B. et PAGES J.
Comparaison de groupes de variables définies sur le même ensemble d'individus. Rapport INRIA Rocquencourt, 1982.

<EsP 84> ESCOFIER B. et PAGES J.
L'Analyse Factorielle Multiple : une méthode de comparaison de groupes de variables : Data analysis and informatics, III. E. Diday et al, North-Holland, 1984.

<Esc 80> ESCOUFFIER Y.
L'analyse conjointe de plusieurs matrices. Biométrie et temps. Société Française de Biométrie, 1980.

<Fou 84> FOUCART Th.
L'analyse factorielle d'opérateurs - Méthodes, programmation et applications : Data analysis and informatics III. E. Diday et al, North-Holland, 1984.

<Gov 83> GOVAERT G.
Classification croisée. Thèse de docteur ès sciences. Université de Paris VI, 1983.

<Gla 81> GLACON F.
Analyse conjointe de plusieurs matrices de données. Thèse de 3ème cycle, Université Scientifique et Médicale de Grenoble, 1981.

<Kob 77> KOBILINSKY A.
Propriétés et utilisation de l'analyse multicanonique par la méthode de Caroll. Analyse des Données et Informatique. INRIA Rocquencourt, 1977.

<Ler 79> LERMANN I.C.
Les présentations factorielles de la classification. RAIRO, Recherche Opérationnelle, VOL 13, n°2, 1979.

<Lhr 76> L'HERMIER DES PLANTES H.
Structuration des tableaux à trois indices de la statistique. Thèse de 3ème cycle, Université de Montpellier, 1976.

<Mas 74> MASSON M.
Processus linéaires - Analyse non linéaires des données. Thèse de doctorat d'Etat - Université de Paris VI, 1974.

<Mod 82> Rapport MODULAD
INRIA, 1982.

<OK 75> OK Y.
Analyse factorielle typologique et lissage typologique. Thèse de 3ème cycle, Université de Paris VI, 1975.

<Ral 79> RALAMBONDRAINY H.
Application de l'analyse multidimensionnelle à l'étude de la charge d'un ordinateur. Thèse de 3ème cycle, Université de Paris VI, 1979.

<Ral 84> RALAMBONDRAINY H.
Le système SICLA : un système intéractif de classification automatique : Data analysis and informatics III - E. Diday et al, North- Holland, 1984.

<Sap 75> SAPORTA G.
Liaisons entre plusieurs ensembles de variables et codage de données qualitatives. Thèse de 3ème cycle, Université de Paris VI, 1975.

<Sap 79> SAPORTA G.
Statistique et Analyse de Données, n°3, 1979, pp 19-31.

<Sch 81> SCHWARTZ L.
Les tenseurs. Editions Hermann, 1981.

<Ten 84> TENENHAUS M.
L'analyse canonique généralisée de variables numériques, nominales ou ordinales par des méthodes de cordage optimal : Data analysis and informatics, III E. Diday et al, North- Holland, 1984.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique